Towards chemometric methodologies on hyperspectral imaging for low dose compound detection : application on Raman microscopy

Mathieu Boiret

Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de

Docteur

Délivré par **l'Université de Montpellier II**

Préparée au sein de l'école doctorale **Sciences des procédés – Sciences des Aliments**
Et de l'unité de recherche ITAP

Spécialité : Génie des procédés

Présentée par **Mathieu Boiret**

TOWARDS CHEMOMETRIC METHODOLOGIES ON HYPERSPECTRAL IMAGING FOR LOW DOSE COMPOUND DETECTION:

APPLICATION ON RAMAN MICROSCOPY

Soutenue le 10 décembre 2015 devant le jury composé de

Mr. Jocelyn CHANUSSOT, Professeur, INP Grenoble	Rapporteur
Mr. Ludovic DUPONCHEL, Professeur, LASIR Lille	Rapporteur
Mr. Jean-Michel ROGER, ICPEF, IRSTEA Montpellier	Directeur
Mme Nathalie GORRETTA, IR, IRSTEA Montpellier	Encadrante
Mr. Serge RUDAZ, Professeur, Université de Genève	Examinateur
Mme Anna de JUAN, Professeur, Université de Barcelone	Invitée
Mr. Douglas RUTLEDGE, Professeur, AgroParisTech, Paris	Invité
Mr. Yves-Michel GINOT, Technologie SERVIER, Orléans	Invité

irstea

SERVIER

« He who knows all the answers

has not been asked all the questions »

Confucius

Remerciements

Ces travaux ont été financés par une convention de recherche entre l'Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture (IRSTEA) et Technologie SERVIER.

Je suis très reconnaissant envers ma hiérarchie, qui m'a donné la chance de vivre cette aventure très enrichissante en parallèle de mes activités au sein de Technologie SERVIER. Ainsi, j'adresse mes remerciements les plus sincères à Patrick Genissel, Directeur de la recherche et biopharmacie (SERVIER), Patrick Wuthrich, Directeur du pôle d'expertise de développement pharmaceutique (SERVIER) et Yves-Michel Ginot, Directeur de la division analytique sur le centre de développement pharmaceutique (SERVIER), pour la confiance qu'ils ont su me donner afin que je puisse relever ce défi.

La direction de cette thèse a été assurée par Jean-Michel Roger, coencadrée par Nathalie Gorretta. Je les remercie tous les deux pour leur pédagogie, leur confiance et pour tous ces moments partagés. Je vous suis très reconnaissant de ces trois années passées à vos côtés.

Je tiens à remercier Ludovic Duponchel et Jocelyn Chanussot d'avoir accepté d'être les relecteurs et évaluateurs de ce manuscrit. Merci également à Serge Rudaz, d'avoir accepté d'intégrer mon jury de thèse. Je remercie très sincèrement Anna de Juan et Douglas Rutledge, membres de mon comité de thèse et invités du jury. Nos collaborations durant ces 3 années ont été pour moi très enrichissantes. Vos remarques et conseils ont été de vraies valeurs ajoutées pour assurer la qualité des travaux et articles. Merci infiniment pour votre temps et votre disponibilité.

Ce travail a été réalisé au sein du département Spectroscopie et Chimiométrie, sur le site SERVIER de développement pharmaceutique d'Orléans. J'ai une pensée pour les membres de cette équipe dynamique (Sylvie, Marc, Frank et Yoann) que j'ai la chance de côtoyer au quotidien et qui a su s'adapter à mes disponibilités parfois limitées. Je ne peux m'empêcher d'avoir une pensée toute particulière pour Loïc Meunier, avec qui l'aventure de ce département a commencé.

J'ai bien évidemment une pensée pour mes parents, qui sont de parfaits exemples de force et de courage face à toutes épreuves, ainsi que pour mes frères à qui j'expliquerai bien volontiers le

fond de ces travaux le moment venu. Il est certain que sans eux, rien de tout cela n'aurait été possible.

Enfin un grand merci à Lesly, qui a su me soutenir et me conseiller dans de nombreuses situations. Merci pour ta patience, ton soutien, pour ce que tu m'apportes depuis le début et pour ce que tu vas m'apporter dans quelques semaines !

Publications and communications

- **Papers in international peer-reviewed journals**

Following articles directly result from this thesis and are referenced in the manuscript as:

Art. I Boiret, M., Rutledge, D. N., Gorretta, N., Ginot Y.M., Roger, J.M. (2014). **Application of independent component analysis on Raman images of a pharmaceutical drug product: Pure spectra determination and spatial distribution of constituents.** Journal of Pharmaceutical and Biomedical Analysis, Vol. 90, 78-84. DOI: 10.1016/j.jpba.2013.11.025

Art. II Boiret, M., de Juan, A., Gorretta, N., Ginot Y.M., Roger, J.M. (2015). **Distribution of a low dose compound within pharmaceutical tablet by using multivariate curve resolution on Raman hyperspectral images.** Journal of Pharmaceutical and Biomedical Analysis, Vol. 103, 35-43. DOI: 10.1016/j.jpba.2014.10.024

Art. III Boiret, M., de Juan, A., Gorretta, N., Ginot Y.M., Roger, J.M. (2015). **Setting local rank constraints by orthogonal projections for image resolution analysis: application to the determination of a low dose compound.** Analytica Chimica Acta, vol. 892, 49-58. DOI: 10.1016/j.aca.2015.08.031

Art. IV Boiret, M., Gorretta, N., Ginot Y.M., Roger, J.M. (2015). **An iterative approach for compound detection in an unknown pharmaceutical drug product: Application on Raman microscopy.** Submitted in Journal of Pharmaceutical and Biomedical Analysis

Other articles:

Art. V Boiret, M., Rutledge, D. N., Gorretta, N., Ginot Y.M., Roger, J.M. (2014). **Raman microscopy and Chemometric tools for counterfeit detection of pharmaceutical tablets.** Spectra Analyse, Vol. 298, 74-80.

- **Oral communications**

Boiret, M., Rutledge, D. N., Gorretta, N., Ginot Y.M., Roger, J.M., **Applications of Raman hyperspectral imaging in the pharmaceutical field**. Invited speaker in "Chimiométrie 2013", 2013, Brest, France.

Boiret, M., **Imagerie chimique par spectroscopies Raman et proche infrarouge : Applications pour l'industrie pharmaceutique**, in "4ème journée de l'expertise chimique à Lyon", 2015, Lyon, France.

Boiret, M., Rutledge, D. N., de Juan, A., Gorretta, N., Roger, J.M., **Use of Chemometric tools on Raman hyperspectral images for low dose constituent distribution in tablets,** in "EuroAnalysis", 2015, Bordeaux, France.

Content

Remerciements iii

Publications and communications v

Content vii

List of figures xiii

List of tables xvii

Abbreviations and notations xviii

Chapter I: General introduction 1

1. Introduction 2

2. Outline of the thesis 3

Chapter II: The use of Raman spectroscopy in the pharmaceutical environment: theory and applications 5

1. Raman spectroscopy 6

1.1. Theoretical aspects 6

1.2. Raman chemical imaging 9

1.3. Applications in the pharmaceutical environment 10

2. Chemometric tools 10

2.1. Data pre-processing 12

2.1.1. Spike correction 12

2.1.2. Baseline correction 13

2.1.3. Normalisation 14

2.1.4. Derivatives 15

2.2. Multivariate data analysis 16

2.2.1. Principal component analysis 16

2.2.2. Independent component analysis 17

2.2.3. Multivariate curve resolution-Alternating least squares 18

3. Identification of a low dose compound 19

3.1. Definition of a low dose compound 19

3.2. The sampling aspect 20

3.3. Data analysis aspect 22

3.4. Contributions of the thesis 23

Chapter III: Use of blind source separation approach for pure spectra determination and spatial distribution of constituents 25

1. Introduction 28

2. Materials and methods 30

2.1. Samples 30

2.2. Raman imaging system 30

2.3. Pre-processing 30

2.4. Independent Component Analysis (ICA) 31

2.5. Data analysis 32

3. Results & discussion 32

3.1. Selection of number of independent components 32

3.2. Distribution of API 33

4. Conclusions 42

Chapter IV: Use of multivariate curve resolution for identification of a low dose compound 48

1. Introduction 51

2. Materials and Methods 53

2.1. Samples 53

2.2. Raman imaging system 53

2.3. Pre-processing 54

2.4. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) 54

3. Results and discussion 56

3.1. Exploratory analysis 56

3.2. MCR-ALS 59

3.2.1. Non-negativity and local rank constraints 59

3.2.2. Effect of PCA filtering on MCR-ALS results 62

3.2.3. Pure spectrum augmented matrix 66

4. Conclusions 68

Chapter V: An alternative method for presence/absence maps determination by orthogonal projections 72

1. Introduction 75

2. Theory 77

2.1. Notations 77

2.2. Pretreatment using orthogonal projections 77

2.3. Multivariate curve resolution-alternating least squares (MCR-ALS) 78

2.4. Proposed approach to determine presence/absence maps of compounds to set local rank constraints 79

3. Materials and methods 82

3.1. Raman microscopy 82

3.2. Samples 82

3.2.1. Simulated data 82

3.2.2. Real dataset 85

4. Results and discussion 85

4.1. Principal component analysis (PCA) on pure images 85

4.2. Proposed approach on simulated data 86

4.3. Proposed approach on real dataset 90

5. Conclusions 93

Chapter VI: An iterative approach for compound detection in an unknown formulation . 95

1. Introduction 97

2. Materials and methods 99

2.1. Notations 99

2.2. Samples 100
2.3. Raman imaging system 100
2.4. Spectral library 100
2.5. Proposed approach 101
2.5.1. Spectral distances 102
2.5.2. Identification of the pure compound 103
2.5.3. Orthogonal projection 104
2.5.4. Overview of the iterative approach 105
3. Results and discussion 106
3.1. Identification of the tablet compounds 106
3.2. Multivariate curve resolution-alternating least squares 114
4. Conclusions 116

Chapter VII: Conclusions and future work 118

1. Introduction 119
2. Main contributions 119
2.1. A flashback to the beginning of this work 119
2.2. Applications of a blind source separation methodology 121
2.3. Applications of multivariate curve resolution 122
2.4. Alternative method for presence/absence map estimations 123
2.5. Compound detection in an unknown formulation 123
3. Limits and future work 124

General conclusion 127

Résumé en français ..... a

1. **Contexte et objectifs** ..... **a**

2. **Matériel et méthodes** ..... **d**

2.1. Instrumentation et échantillons ..... d

2.2. Analyse des données ..... d

3. **Contributions** ..... **e**

4. **Résultats** ..... **f**

4.1. Utilisation de la séparation de source aveugle pour la détermination de spectres purs et l'étude de la distribution spatiale des composés ..... f

4.2. Utilisation de résolution multivariée de courbes pour l'identification d'un constituant faiblement dosé ..... h

4.3. Proposition d'une méthode pour la mise au point des cartographies d'absence/présence de composés ..... k

4.4. Approche itérative pour la détection des composés d'une formulation inconnue ..... m

5. **Conclusions** ..... **p**

List of figures

Figure II-1 Description of the Raman scattering....8

Figure II-2 Generation of hyperspectral data cube....9

Figure II-3 Application of spike correction on a lactose spectrum....13

Figure II-4 Application of baseline correction using Asymmetric Least Squares on 25 spectra of microcrystalline cellulose....14

Figure II-5 Application of SNV correction on 25 spectra of Amlodipine....15

Figure II-6 Example of derivative correction on 25 spectra of aspartame....16

Figure II-7 – Probability of finding at least one spectrum of a 0.5% w/w low dose compound....21

Figure III-1 Graphical representation of the tested approach....26

Figure III-2 Lowest correlation between signals obtained using ICA_by_blocks. The lowest correlation obtained using the ICA_by_blocks approach significantly decreases after 9 ICs, which was considered as the optimal number of component for the decomposition of the dataset....33

Figure III-3 Proportions coefficients (A) of each IC. Images correspond to the proportions coefficients (A) of a 9 ICs model. A red color corresponds to a high value whereas a blue color corresponds to a low value.....34

Figure III-4 Signals, S, of the ICA model. These signals correspond to the calculated signals (S) of a 9 ICs model.....35

Figure III-5 Pure spectra of the drug product constituents. In blue API 1, in green API 2, in black lactose, in red avicel and in magenta the magnesium stearate. Relative intensities were used as the spectra were split for a better observation.....36

Figure III-6 Calculated signal of independent component 9 superposed on the spectrum of API 1. Comparison between API 1 spectrum and IC9 signal. The correlation between the two signals is equal to 0.92.....37

Figure III-7 Calculated signal of independent component 6 superposed on the spectrum of API 2. Comparison between API 2 spectrum and IC6 signal. The correlation between the two signals is equal to 0.96. 38

Figure III-8 Calculated signal of independent component 2, 3, 4, 5 plotted with the spectrum of Lactose. Comparison between lactose spectrum and IC2, IC3, IC4 and IC5 signals. The correlations between the signals are respectively equal to 0.44, 0.23, 0.25 and 0.47. The pure spectrum of lactose and the four calculated independent components are displayed. The pure spectrum was decomposed into four components. 39

Figure III-9 IC3 signal from a 5 components ICA model superposed on the spectrum of lactose. Comparison between lactose spectrum and IC3 signal from a 5 components ICA model. The correlation between the two signals is equal to 0.90. 40

Figure III-10 IC12 superposed on the magnesium stearate spectrum from a 15 component ICA model. Comparison between magnesium stearate spectrum and IC12 signal from a 15 components ICA model. The correlation between the two signals is equal to 0.87. 41

Figure III-11 Distribution of IC12 (magnesium stearate) from a 15 component ICA model. This component is highly correlated to magnesium stearate. 42

Figure III-12 Article V: Application of independent component analysis on counterfeit samples 47

Figure IV-1 General scheme of the tested approaches in Chapter IV 50

Figure IV-2 Preprocessed Raman spectra (AsLS and first derivative) 57

Figure IV-3 PCA scores: five first components associated with their explained variances. Different distributions and agglomerates were highlighted. PC1 and PC5 were linked to the lactose variability, while PC2, PC3 and PC4 were respectively linked to the distributions of API1, avicel and API2. 58

Figure IV-4 Singular values plot (top: non-sorted singular values, bottom: sorted singular values) 60

Figure IV-5 Local rank map obtained by choosing an appropriate threshold which separates significant singular values from noise. 61

Figure IV-6 C_{sel} matrix (Orange: absence of the constituent, White: presence of the constituent) 61

Figure IV-7 Highest correlation between the calculated spectra (S_{opt}) and the reference spectrum of magnesium stearate (for each iteration of a PCA filtered matrix built from 5 to 100 components) 63

Figure IV-8 Distribution maps of drug substance constituents (PCA non-filtered dataset) 65

Figure IV-9 S_{opt} versus reference spectrum of magnesium stearate 65

Figure IV-10 Distribution maps of drug substance constituents (augmented matrix approach) . 67

Figure V-1 Graphical representation of the proposed approach 81

Figure V-2 Image distribution patterns used to build synthetic image (eight classes represented by eight different colours) 83

Figure V-3 Simulated distribution maps of lactose, avicel®, API and magnesium stearate (MgSt) 84

Figure V-4 Building of simulated data 85

Figure V-5 Non-centered PCA on pure compound image of API. Raw spectra, scores maps and loadings 86

Figure V-6 Orthogonal projected spectra of the simulated data to the suitable interference space K_{-c} for each compound 87

Figure V-7 Correlation maps $r(x_{n\perp}, s_{\perp})_i$ (k1 = lactose basis, k2 = avicel basis, k3 = API basis, k4 = MgSt basis) 88

Figure V-8 Presence/Absence maps of compounds in the simulated image (blue colour: Presence of the compound, white colour: absence of the compound) 89

Figure V-9 Presence/absence maps of drug compounds (blue colour: Presence of the compound, white colour: absence of the compound) 90

Figure V-10 Calculated spectrum by MCR-ALS (with n-n and local rank constraints) and pure spectrum of magnesium stearate 91

Figure V-11 Distribution maps of the five compounds obtained by MCR-ALS (with n-n and local rank constraints) 92

Figure VI-1 – Description of α angle in the spectral angle mapper (SAM) calculation 103

Figure VI-2 – Boxplot representation of SAM 104

Figure VI-3 – Description of the proposed approach 106

Figure VI-4 – Raw spectra of the image of dimensions 30 pixels per 30 pixels (900 spectra) 107

Figure VI-5 – Mean spectra of the 24 pure products included in the spectral library 108

Figure VI-6 – SAM values boxplot and $mq1$ values calculated from iteration 1 to 8 (to identify each sample number, readers are refered to Table VI-1) 110

Figure VI-7 – SAM values between pure projected spectrum RMgSt ⊥of magnesium stearate and the X ⊥ matrix (iteration 7) 111

Figure VI-8 – Projected spectrum $X_{\perp}$ at positions y = 23 and x = 4 and the projected mean spectrum $RMgSt_{\perp}$ 112

Figure VI-9 – Evolution of standard deviation of the 24 $mq1$ values 113

Figure VI-10 – Distribution maps of metolose, API form 1, eudragit, microcrystalline cellulose, API form 2, magnesium stearate and maltodextrin 115

List of tables

Table II-1 Spatial and spectral contributions of a compound 19

Table III-1 Correlation coefficients between the ICA signals and the pre-processed true compound spectra. The comparison between the calculated signals and the true spectrum of each compound shows that only two ICs are directly linked to the drug product constituents. For each component, the highest correlation was highlighted with bold characters..... 36

Table IV-1 Correlations between MCR-ALS calculated S_{opt} and the reference spectra (PCA filtered dataset) 62

Table IV-2 MCR-ALS results according to the number of components used to build the PCA reduced DPCA(n, p) matrix..... 64

Table IV-3 Correlations between MCR-ALS S_{opt} and the reference spectra (column-wise augmented dataset) 67

Table V-1 Target concentrations of the eight classes 83

Table V-2 MCR-ALS results on simulated data 89

Table V-3 MCR-ALS results on real dataset..... 92

Table VI-1 – Spectral library..... 101

Table VI-2 – Estimation of the compound concentration 115

Abbreviations and notations

Abbreviations

API	Active Pharmaceutical Ingredient
AsLS	Asymmetric Least Squares
BSS	Blind Source Separation
CLS	Classical Least Squares
ICA	Independent Component Analysis
IR	Infrared
JADE	Joint Approximate Diagonalization of Eigenmatrices
lof	Lack of fit
MCR	Multivariate Curve Resolution
MCR-ALS	Multivariate Curve Resolution-Alternating Least Squares
NAS	Net Analyte Signal
NIR	Near infrared
OPA	Orthogonal Projection Approach
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
PLS-DA	Partial Least Squares-Discriminant Analysis
QC	Quality Control
R2	Explained variance
SIMPLISMA	Simple-to-use interactive Self-modeling Mixture Analysis

Notations

$\mathbf{x}$	Vectors in bold lowercase
$\mathbf{X}$	Matrix in bold uppercase
$\bar{\mathbf{x}}$	Mean vector of $\mathbf{X}$
n	Scalar in italic lowercase
$\mathbf{X}^T$ and $\mathbf{x}^T$	Transpose form of a matrix $\mathbf{X}$ and a vector $\mathbf{x}$
$\mathbf{K}$	Spectral basis, vector subspace
$\mathbf{\Sigma}$	Euclidian projector to a spectral basis $\mathbf{K}$
$\mathbf{x}_\perp$ and $\mathbf{X}_\perp$	$\mathbf{X}$ and $\mathbf{x}$ orthogonally projected to $\mathbf{K}$
$\mathbf{I}$	Identity matrix of dimensions p x p (p variables)
$\bar{\nu}$	Raman shift in cm^{-1}
λ	Wavelength
h	Planck constant
$\lambda_{incident}$	Wavelength of the incident photons
$\lambda_{scattered}$	Wavelength of the scattered photons
N_n	Number of molecules in the excited state
N_m	Number of molecules in the ground state
k	Boltzmann constant
T	Temperature
ΔE	Energy differences between the vibrational energy stated
g_n	Degeneracies of the excited state
g_m	Degeneracies of the ground state
Var_i	Theoretical spectral variance of a compound i
$\mathbf{C_{sel}}$	Absence/presence matrix
$\mathbb{R}^n$	N-dimensional space in which variables can be represented as vectors
$\mathbb{R}^p$	P-dimensional space in which observations can be represented as vectors

Chapter I: General introduction

1. **Introduction** 2

2. **Outline of the thesis** 3

1. Introduction

In the pharmaceutical environment, and especially in the research and development field, the quality of the medicine is a critical step as it is facing challenges with increased demand from the regulatory affairs to improve the quality of a pharmaceutical drug product. In order to ensure its proper effect on the patient health, a product has to be manufactured with the appropriate quality [1-3].

Today, a lot of techniques are used in quality control (QC) laboratories to ensure the quality of a drug product. Several tests such as dissolution profiles, stability studies or control of active content are required from the pharmaceutical guidelines and authorities to ensure that the analysed product is included within pre-determined specifications. In the QC labs, most of the analytical tools are based on chemical analyses (liquid chromatography, dissolution apparatus…) which generally damage the sample, require solvent and a lot of time or important human resources.

In the last decade, the use of vibrational spectroscopy such as near infrared or Raman spectroscopy has grown quickly and has appeared as an alternative analytical tool to usual techniques [4; 5]. By allowing fast and non-destructive analysis, without needing sample preparations in most cases, these analytical tools are particularly appreciated by the analysts. New available guidelines from European Pharmacopeia [6] or European Medicine Agency (EMA) [7] have strongly encouraged the use of these alternative techniques in the QC laboratories. The main objective is to continuously improve the knowledge of a pharmaceutical drug product to produce a medicine with high and consistent quality [8].

Due to the complexity of the acquired spectra or because univariate observation of the data can be inadequate, multivariate data analysis and chemometrics are often needed to extract useful information from spectroscopic measurements [9]. Several applications have been previously published in the pharmaceutical environment. Qualitative analysis such as raw material identification [10] or counterfeit detection [11], have been carried out and have particularly been appreciated in the pharmaceutical field. Quantitative methods such as content uniformity, quantification of a crystalline form during stability studies, have been developed in order to replace usual chemical approaches [12].

Apparition of chemical imaging, which gives both spectral and spatial information on the studied sample by associating two spatial dimensions (x and y dimensions) and one spectral dimension (each pixel spectrum) provides a new way of exploring a sample, i.e. a pharmaceutical drug

product [13; 14]. Indeed, by adding the spatial information, it is now possible to study the distribution of actives and excipients within a tablet or a powder sample. In the case of Raman microscopy, hyperspectral imaging (also called chemical imaging) is the association of a microscope and a Raman spectrometer. Because of the huge amount of data contained in hyperspectral images, a direct interpretation of the acquired images is not possible. Therefore, several chemometric tools have previously been applied for qualitative or quantitative analysis of hyperspectral dataset [15]. Some methods are mainly based on variance decomposition, while other methods require a calibration step or prior knowledge to develop predictive models. In most pharmaceutical cases, Raman microscopy coupled with chemometrics was used to study the compound distributions in a sample. Indeed, the study of active and excipient distributions can be viewed as a critical parameter significantly influencing the quality of the tablet. A non-controlled distribution can have an impact on the tablet dissolution profile or can facilitate the apparition of degradation products which may be one of the reasons of a troubleshooting alert throughout the manufacturing process.

In the framework of compound distributions, the study of a low dose compound, which can be viewed as a product located in a few pixels of an image and with a low spectral contribution comparing with other products, appeared as a real challenge. Indeed, because information linked to this product is weak and because chemometric algorithms are mainly based on the decomposition of statistical moments, detection of a low dose product could be difficult.

The main objective of this thesis will be to study the ability of different chemometric tools and methods:

i/ to study the compound distributions within a pharmaceutical drug product

ii/ to identify a low dose compound in a pharmaceutical drug product

To reach these objectives, different chemometric tools will be tested, with or without prior knowledge on the formulation, and innovative methodologies will be proposed, developed and applied on simulated and real case Raman hyperspectral datasets.

2. Outline of the thesis

The thesis consists of an introductory part (chapter I), followed by a state of the art section (chapter II) on the use of Raman spectroscopy in the pharmaceutical environment. In these two first sections, the major aspects of Raman spectroscopy and chemometrics tools are presented

with the help of several applications in the pharmaceutical environment. Brief introduction of the Raman effect, spectral interpretation and data analysis will be described based on a review of Raman applications in the pharmaceutical field. Moreover, the main objective of this thesis, which can be resumed as the detection of a low dose compound within a pharmaceutical drug product, will be detailed and explained.

In the following sections, each chapter of the thesis refers to a scientific publication (published or submitted), forming the spine of this manuscript. The author would like to apologize for any potential redundancies between the different chapters, especially in the materials and methods sections, due to the chosen format of the thesis, based on articles.

Chapter III discusses the ability of a blind source separation method, independent component analysis, to extract pure compound signals in hyperspectral dataset without prior knowledge. This chapter is the reproduction of **Art. I** published in the Journal of Pharmaceutical and Biomedical Analysis in 2014.

Chapter IV details the use of multivariate curve resolution-alternating least squares to resolve a system including a low dose compound. Different approaches will be tested and discussed in this section. This chapter is the reproduction of **Art. II** published in the Journal of Pharmaceutical and Biomedical Analysis in 2015.

In Chapter V, an innovative procedure to set the presence/absence maps of compounds for later use as local rank constraints in the multivariate curve resolution-alternating least squares iterative process is proposed. The algorithm is based on orthogonal projection to a space containing the contributions to be removed (i.e. the interference subspace) and spectral comparison between the projected spectrum and a pure projected spectrum of the compound of interest. This chapter is the reproduction of **Art. III** published in Analytica Chimica Acta in 2015.

In chapter VI, an iterative approach is proposed to identify the pure compounds of a unknown pharmaceutical drug product by using a spectral library, spectral distances and orthogonal projections. This chapter is the reproduction of **Art. IV** submitted in the Journal of Pharmaceutical and Biomedical Analysis in 2015.

Finally, the last chapter of this thesis (chapter VII) concludes by synthesizing the key points of the tested and developed approaches. It proposes some perspectives and future research applications to continue this work.

Chapter II: The use of Raman spectroscopy in the pharmaceutical environment: theory and applications

1. Raman spectroscopy 6

1.1. Theoretical aspects 6

1.2. Raman chemical imaging 9

1.3. Applications in the pharmaceutical environment 10

2. Chemometric tools 10

2.1. Data pre-processing 12

2.1.1. Spike correction 12

2.1.2. Baseline correction 13

2.1.3. Normalisation 14

2.1.4. Derivatives 15

2.2. Multivariate data analysis 16

2.2.1. Principal component analysis 16

2.2.2. Independent component analysis 17

2.2.3. Multivariate curve resolution-Alternating least squares 18

3. Identification of a low dose compound 19

3.1. Definition of a low dose compound 19

3.2. The sampling aspect 20

3.3. Data analysis aspect 22

3.4. Contributions of the thesis 23

1. Raman spectroscopy

1.1. Theoretical aspects

The objective of this chapter is to provide a brief introduction to Raman spectroscopy for people that are not familiar with this technology. For more details, readers are referred to the literature [16-18].

Raman spectroscopy can be considered as a vibrational spectroscopy. When an electromagnetic wave interacts with electrical and magnetic fields of atoms or molecules, different phenomena are observed depending on the energy of said wave. Optical spectroscopies constitute the body of methods that measure these light/matter interaction phenomena and thus use light for the study of molecular processes. Vibrational spectroscopies are optical spectroscopy techniques based on transitions between vibrational levels of the same electronic state. It measures the interaction of the incident electromagnetic radiation with the specific molecular vibrations of the sample. From spectrum, it is thus possible to deduce information on the nature and structure of a molecule, in either free or bonded form, as well as its interaction with its environment [19; 20].

Raman spectroscopy uses a monochromatic light source (typically a laser). When light (of frequency ν_0) interacts with matter, incident photons are mainly transmitted and absorbed by the sample molecules. However, a slight part of the incident light is also scattered. In that case, most of the photons (1 photon / 10^4 photons) are elastically scattered, meaning that they have the same energy of the incident light. This phenomenon is called the Rayleigh scattering effect. Occasionally (1 photon / 10^8 photons), a photon can be "inelastically" scattered, meaning that it has a frequency different that the frequencies of the incident light. This phenomenon corresponds to the Raman effect (Figure II-1). If the frequency of the scattered light is lower than the frequency of the incident light, then the Stokes Raman effect is measured ($h\nu_S = h\nu_0 - h\nu_v$). However, if the frequency of the scattered light is higher than the frequency of the incident light, then the anti-Stokes effect is measured ($h\nu_{AS} = h\nu_0 + h\nu_v$). The Raman effect is weak comparing with the Rayleigh effect. In experimental applications, the Raman Stokes scattering is mainly measured as its intensity is higher than the anti-Stokes effect [21; 22]. Indeed, because the majority of molecules are in the ground energy state at room temperature, and not in an excited state as required for generating anti-Stokes scattering, the Stokes scattering is mainly observed. This observation can be explained by the Boltzmann distribution which describes the relationship between temperature and the fraction of molecules in an excited state [23]:

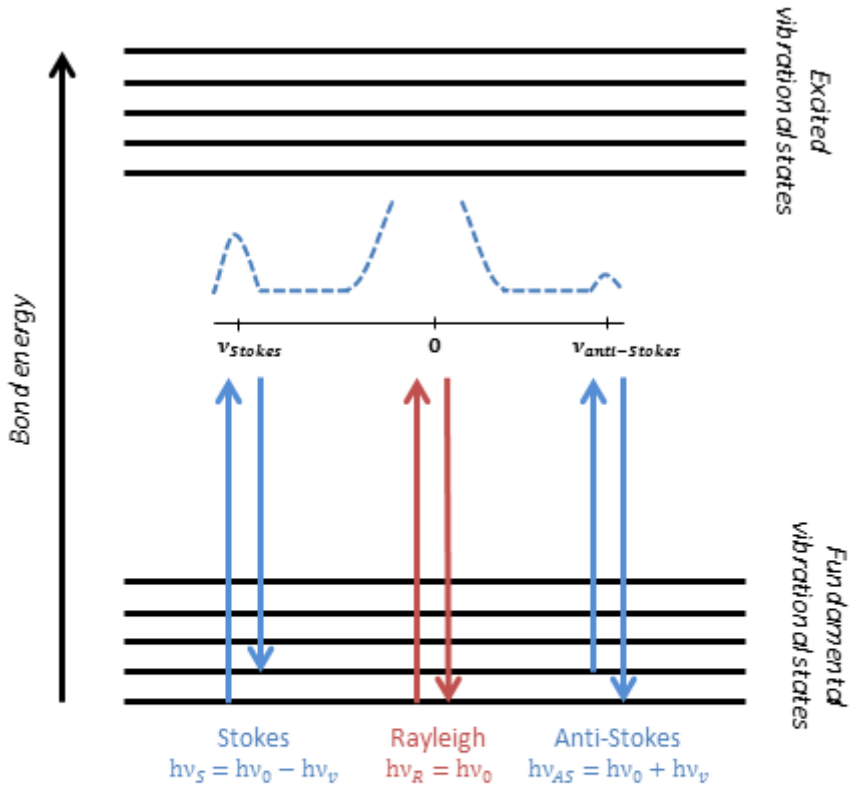
$$\frac{N_n}{N_m} = \frac{g_n}{g_m} exp^{\frac{-\Delta E}{kT}} \qquad \textbf{(II-1)}$$

where N_n and N_m are the number of molecules in the excited and ground energy states, g_n and g_m the degeneracies of the excited and ground vibrational states, k the Boltzmann constant, T the temperature in Kelvin and ΔE the energy differences between the vibrational energy states. If the temperature increases, the number of molecules in the excited state increases and the anti-Stokes intensity changes accordingly. In theory, Stokes and anti-Stokes measurements contains the same frequency information, with different intensity levels.

A Raman spectrum represents the intensity of Stokes or anti-Stokes lines as a function of wavenumber and not frequency. The x-axis is generally labelled the Raman shift ($\bar{\nu}$) and measured in wavenumber (cm^{-1}). It can be calculated using the following equation:

$$\bar{\nu} = \left(\frac{1}{\lambda_{incident}} - \frac{1}{\lambda_{scattered}}\right) \times 10^7 \qquad \textbf{(II-2)}$$

Where $\lambda_{incident}$ and $\lambda_{scattered}$ are the wavelengths of the incident and the Raman scattered photons [24]. The Raman shift is then independent of the incident light frequency (the characteristic bands on a Raman spectrum will be the same whatever the wavelength of the laser). The positions of the Raman shifted wavenumbers for a given vibrational mode are identical to the wavenumbers of the corresponding bands in an infrared absorption spectrum. However, the stronger peaks in a Raman spectrum are often weak in an infrared spectrum, and vice versa. Comparing with other analytical tools, Raman spectroscopy is advantageous because quick and accurate measurements can often be made without destructing the sample and with minimal or no sample preparation.



h : Planck's constant
ν_0 : Excitation frequency
ν_v : Vibration frequency of the bond
ν_S : Frequency of the Stokes line
ν_{AS}: Frequency of the anti-Stokes line

Figure II-1 Description of the Raman scattering

A lot of Raman instrumentations are available on the market and the objective of this chapter is not to provide an exhaustive review of the Raman technologies. However, Raman apparatus can be briefly described as a system constituted of a monochromatic light source, a filter to remove the Rayleigh scatter, a spectrograph to separate the Raman scattered light by wavelength, a detector, and a computer to visualize the data [25]. Different light sources (lasers) are available. The choice of excitation wavelength is a compromise: the higher the energy of the wave, the more intense scattering but also the greater risk of inducing parasitic fluorescence. In the pharmaceutical environment, to study tablet or power, a 785nm laser is often a judicious choice.

1.2. Raman chemical imaging

Hyperspectral Raman images can be acquired by using two different modes: Raman mapping or Raman imaging. In the first case, the spectrum of the sample is dispersed across the detector, and the sample is moved when each spectrum has been measured. In the second case, the image of the sample at a single wavelength is focused on the detector and the wavelength is changed after each measurement [26].

In most pharmaceutical applications, Raman mapping systems are used. Raman spectroscopy is coupled with a microscope in order to acquire both spectral and spatial information of a sample. The Raman images make possible the characterisation of the pure compound in a pharmaceutical drug product and can provide the distribution of actives and excipients on the surface of a sample. The acquisition system generates hyperspectral data cube, defined by the spatial dimensions *x* and *y* and the spectral dimension *p*, corresponding to the Raman shift (Figure II-2). The easiest way of having an image is to observe the data at a specific Raman shift, but this visualisation can only be carried out when specific Raman bands are available. It is not suitable when signals are overlapped. In practice, spectral datasets are often composed by hundreds of variables (i.e. Raman shift) which makes the direct visualisation difficult. Due to high correlations between variables, their dimensions can be (mathematically) reduced without losing a lot of information [27].

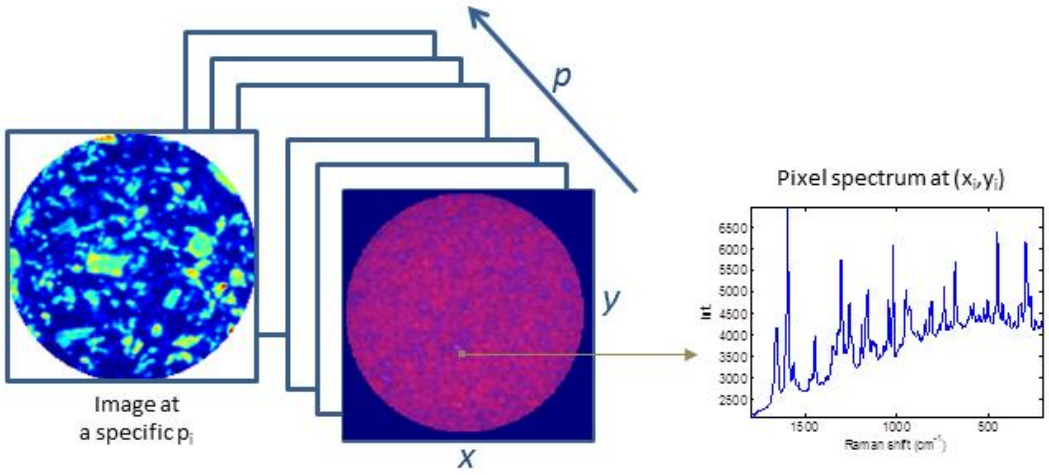


Figure II-2 Generation of hyperspectral data cube

Every pixel contains a Raman spectrum, which can be a mixture of different compounds, depending on the spatial resolution used, which is a critical parameter of the analysis. Regarding to the Raman system, it can vary from hundreds of nanometres (high spatial resolution) to

hundreds of micrometres (low spatial resolution). The lower the spatial resolution, the greater risks of acquiring mixture signals for each pixel. However, the higher the spatial resolution, the longer acquisition time. A compromise must be selected depending on the objective and the acquisition time.

1.3. Applications in the pharmaceutical environment

In the pharmaceutical environment, Raman spectroscopy is used for qualitative or quantitative analysis [28; 29]. The continuous improvement and simplification of apparatus have made the use of this analytical tool easier for analysts without knowing the Raman theory in details. The rapid, non-destructive and non-invasive features of this technology mark its potential suitability as a process analytical tool for the pharmaceutical industry, for both process monitoring and quality control throughout drug production [30]. Chemical imaging can be considered as an emerging platform technology that integrates conventional imaging and spectroscopy to attain both spatial and spectral information from a sample.

In the pharmaceutical environment, Raman spectroscopy and chemical imaging have been previously used in various ways [31]:

- Raw material identification in warehouses [10]
- Quantitative determination of active substance in a solid drug product [32; 33]
- Detection and quantification of crystalline forms [34; 35]
- Fight against illegal drugs / Counterfeit detection [36-38]
- Process Analytical Technology: Support chemical or pharmaceutical development [39-41]
- Pharmaceutical development: determination of the tablet homogeneity [42; 43], understand dissolution performance [44]…

2. Chemometric tools

A Raman spectrum contains a lot of information that describes the chemical and physical composition of a sample. In the case of chemical imaging, the amount of data can be very important and a visual interpretation of the data is not possible. In order to extract the useful information, two approaches can be considered: univariate and multivariate data analysis. Historically, analysis of Raman data has been limited to the univariate approach by analysing

Raman band intensities or by calculating Raman band ratio. Univariate analysis is considered as the easiest, most prevalent and most robust data analysis approach and, in many cases, can provide sufficient information and reliable predictability [45]. But, most of the time, the complexity and the amount of data require the use of multivariate data analysis.

In most cases, chemometric tools appeared as a powerful solution to extract the desired information. Indeed, chemometrics was extremely useful to investigate complex and very similar spectra by extracting the relevant chemical information from the raw spectra, especially when they have a large number of variables and significant overlap of analytical signals. Another advantage is that chemometric tools are statistical methods which provide an objective way to examine spectra, as opposed to pure visual inspection [46]. Large datasets are generated using Raman spectroscopy, thus, extracting targeted information from these complex datasets is a real challenge.

Several chemometric methods have been developed and applied on spectroscopic data and hyperspectral imaging [15; 47]. In most applications of chemical imaging, data analysis procedure consists of the following steps [48] :

- Unfold the image (3-dimensions dataset to 2-dimensions dataset)
- Pre-process the data (spike, baseline correction…)
- Perform data analysis (unsupervised or supervised algorithms)
- Fold results back to image (distribution maps)
- Enhance resulting image (image filtering, contrast enhancement…)

In order to apply conventional chemometric tools on hyperspectral data cube, chemical images are usually unfolded from a 3-dimensions dataset to a 2-dimensions dataset. Common chemometric tools can then be applied on the unfolded hyperspectral images. By using these techniques, all the spectral information of the data cube is taken into account. Principal component analysis (PCA) [49], classical least squares (CLS) [50], partial least squares (PLS) [51], multivariate curve resolution (MCR) [52], partial least squares-discriminant analysis (PLS-DA) [53] or independent component analysis (ICA) [54] have been previously applied on hyperspectral dataset acquired by vibrational spectroscopy. Three of them were mainly used in this thesis: principal component analysis, independent component analysis and multivariate curve resolution-alternating least squares. By folding back the results, distribution maps can be obtained. Mathematical treatments or filters can be applied on the distribution maps in order to enhance image contrast, or to smooth the image, or to enhance edges in the image [55].

Some of the basic principles of these chemometric methods will be explained in the next sections as well as the pre-processing step of the data.

2.1. Data pre-processing

Pre-processing of the data is often necessary before applying chemometric methods in order to improve the model performance by removing perturbing effect or to enhance slight variations in the dataset [56]. With vibrational spectroscopy, it can be very important to decrease the influence of various signal sources that are not related to the useful chemical or physical information. Light scattering, variations during long acquisitions or different particle-size distributions could have a huge impact on the spectral quality and the use of pre-processing tools is often required.

A lot of pre-processing methods have been previously used on vibrational spectroscopy [57; 58]. Centering of the data, baseline correction or normalisation methods are very famous pre-processing techniques with vibrational spectroscopy. The use of derivative methods, coupled with a smoothing step (such as the Savitzky-Golay algorithm [59]) can be useful to enhance slight variations in the spectral dataset. In the case of Raman spectroscopy, cosmic rays can be observed on spectra, thus spike correction can also be necessary. Even if the objective of this thesis is not to provide an in-depth description of all the pre-processing tools applied on Raman spectroscopy, a brief description of the main approaches cited in this manuscript is provided in the next sections.

2.1.1. Spike correction

Spikes are usually sharp Raman bands which can influence the variance structure of the dataset. There are mainly explained by cosmic rays and high energy particles, striking the CCD (charge-coupled device) detector. These cosmic rays must be removed without modifying the Raman spectral bands before applying chemometrics tools. A lot of methods are available to correct these artefacts [60] and some of them have been successfully applied to Raman chemical imaging [61; 62]. In this work, a spatial approach based on [63] is applied. Image spectra in a square pixel area neighbourhood are used to identify outlier-contaminated data points in the central pixel of that neighbourhood. A preliminary “despiking” of the neighbouring spectra is performed by median filtering. Correlations between the central pixel spectrum and its “despiked” neighbours are calculated, and the most highly correlated spectrum is used to identify outliers. Spike-contaminated data are replaced using results of polynomial interpolation. Application of the spike correction is illustrated in Figure II-3, where signals

before and after the spike correction of a lactose spectrum are displayed. The spike at 610 cm^{-1} was clearly eliminated by the algorithm without modifying the Raman spectrum.

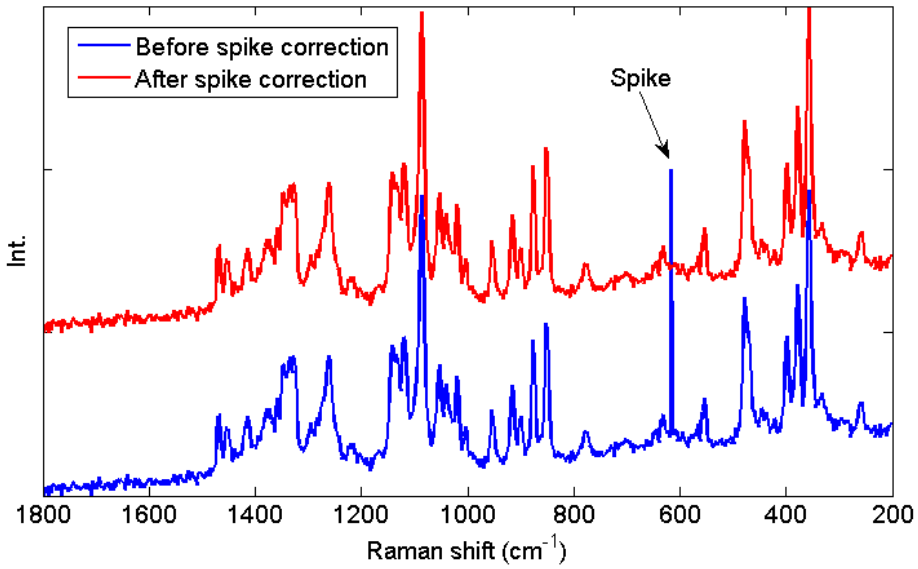


Figure II-3 Application of spike correction on a lactose spectrum

2.1.2. Baseline correction

A spectrum can be considered as the sum of a signal which contains the useful information on the chemical composition of a sample and a background signal which corresponds to the harmful information. With Raman spectroscopy, baseline variations can mainly arise from fluorescence effect by causing the disappearance of the Raman bands. Manual [64], semi-automated [65; 66] or fully automated methods [67] can be applied on the data to correct these unwanted spectral variations.

In [66], a semi-automated method for fluorescence subtraction, based on a modification to least-squares polynomial curve fitting was described. The method was improved in [65] with the addition of a peak-removal procedure during the first iteration and a statistical method to account for signal to noise effects. Experimental results demonstrate that this approach improves the rejection of the fluorescence background during real-time Raman spectroscopy and for in vivo measurements characterized by low signal-to-noise ratios. To avoid the use of parameters such as the polynomial order selection, fully automated baseline correction techniques were developed [67].

Asymmetric least squares (AsLS) is also a powerful method for removing baseline offset from raw Raman spectra. With this approach, it is assumed that some variables contain only background contributions. A polynomial is fitted to each spectrum and variables below the polynomial are up weighted before the next iteration. Process is repeated until that a pre-defined number of variables is reached [68-70]. Application of AsLS pre-processing step was illustrated in Figure II-4 on 25 Raman spectra of microcrystalline cellulose. Comparing with the raw spectra, the baseline variation was successfully corrected by the algorithm.

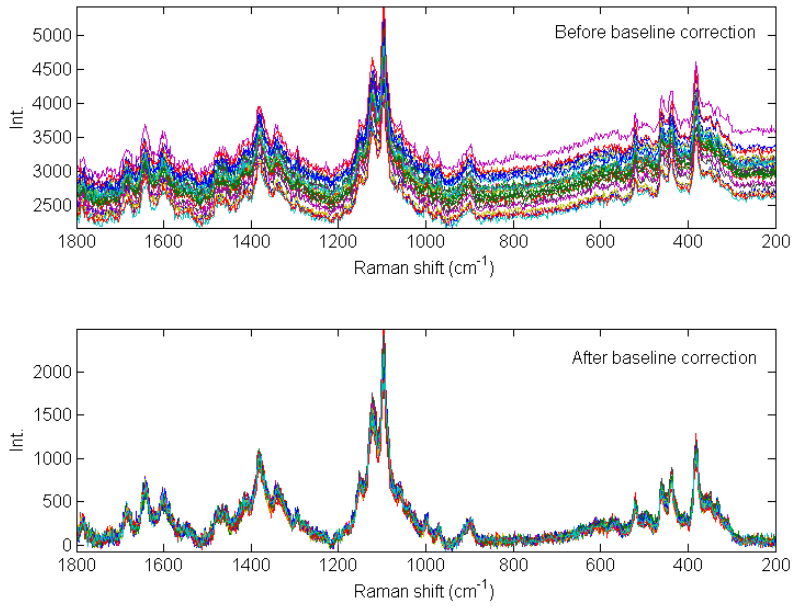


Figure II-4 Application of baseline correction using Asymmetric Least Squares on 25 spectra of microcrystalline cellulose

2.1.3. Normalisation

Due to acquisition variability and to concentrations or scattering variations of a compound, In some cases, Raman intensities can be different between samples or during the whole acquisition. In most situations, a normalisation has to be applied by dividing each variable of a spectrum with a constant [71]. The constant can be the maximum value of a spectrum, or the sum of all variables from a spectrum (also called the normalisation to unit area), or the sum of squares of all variables from a spectrum (also called the normalisation to unit length). Other methods such as standard normal variate (SNV) [72] or multiplicative scatter correction (MSC) [73], previously applied on near infrared spectra, have also been used successfully on Raman dataset

[74]. In Figure II-5, the SNV pre-processing was applied on 25 Raman spectra of Amlodipine. Spectral variability observed in the raw spectra was successfully corrected.

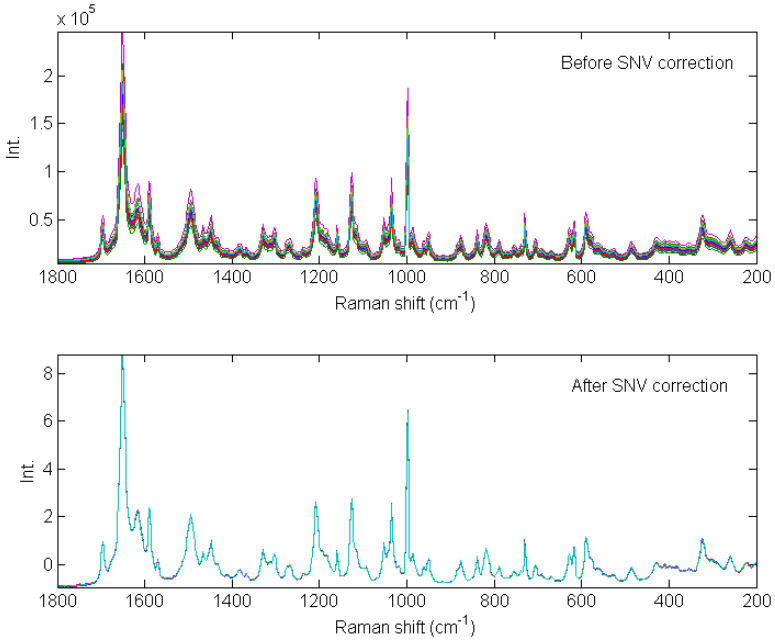


Figure II-5 Application of SNV correction on 25 spectra of Amlodipine

2.1.4. Derivatives

Derivatives can be applied on spectral data for two objectives. The first one is the correction of the baseline variations and the second one is the enhancement of the slight spectral variations. Most applications used a Savitzky-Golay [59] derivation which combines a smoothing and a derivative steps. With derivatives, the signal quality can decrease because the noise will be enhanced. A well-defined compromise has to be chosen between the derivative order, the polynomial order and the window size in accordance with the expected spectral quality.

In Figure II-6, a second order derivative with a window size equal to 9 and 2nd polynomial order, was applied on 25 Raman spectra of aspartame. Baseline variations were significantly decreased and slight spectral variations were enhanced by preserving a sufficient spectral quality.

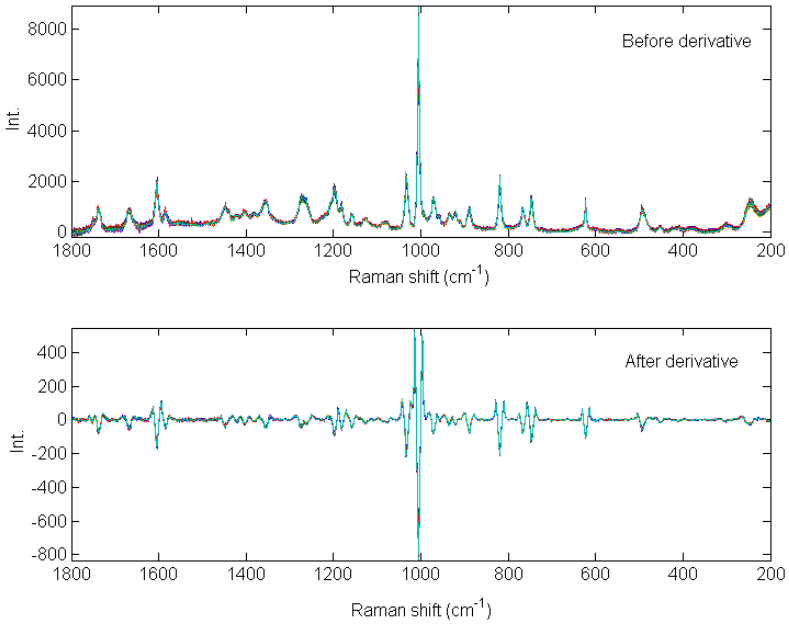


Figure II-6 Example of derivative correction on 25 spectra of aspartame

2.2. Multivariate data analysis

2.2.1. Principal component analysis

The main goal of principal component analysis (PCA) is to reduce the dimensionality of a matrix by removing correlations between variables. PCA decomposes the data in a new set of variables called principal components progressively explaining the largest variations of the dataset [75]. The second principal component is orthogonal to the first one and explains the residual variance not taken into account by the previous one. A spectral matrix **X** can be explained by the score matrix **T**, a loading matrix **P** and a residual matrix **E** with the equation:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \qquad \textbf{(II-3)}$$

Scores refer to spectral variations while loadings represent the spectral contributions to each principal component. For a specific number of components, the residual matrix contains the non-explained information. It will decrease with the number of principal components. With chemical imaging applications, score results can be folded back on order to observe the pixel variability for each principal component. In the case of Raman spectra, where the variables are highly correlated, the number of components is usually considerably lower than the number of variables. PCA can be viewed as a specific case of eigen-decomposition on the variance-

covariance spectral matrix [76]. It is considered as a very powerful tool for exploratory analysis or dimension reduction, and it can also be an interesting tool to detect the number of components in a mixture dataset.

PCA was successfully applied on a lot of vibrational datasets and hyperspectral images [77; 78]. In some cases, the variability associated to a principal component can be linked with a chemical compound of a tablet but in most applications, due to their unclear chemical meaning, loadings and associated images are difficult to interpret.

2.2.2. Independent component analysis

Independent component analysis (ICA) is one of the most powerful techniques in blind source separation [79; 80], assuming that each row of the studied matrix is a weighted sum of pure source signals. It has been developed to extract the pure underlying signals from a set of mixed signals in unknown proportions. Considering a noise-free ICA model, a matrix **X** (*n* x *m*) is decomposed as a linear generative model by the following expression:

$$\mathbf{X} = \mathbf{AS} \qquad \textbf{(II-4)}$$

Where **S** is a (*k* x *m*) matrix of *k* independent source signals called the independent components and **A** is a (*n* x *k*) mixing matrix of coefficients or proportions of the pure signals in each mixed signal of **X**. The objective of ICA is to estimate a set of vectors that are as independent as possible, and the mixed signals in **X** can then be expressed as linear combinations of these independent components (ICs). It attempts to recover the original signals by estimating a linear transformation, using a criterion which reflects the statistical independence among the sources.

To solve the previous equation, an unmixing matrix **W** based on the observation of **X** needs to be calculated. The output **U**, constituted by the independent component $\mathbf{u_1}$,$\mathbf{u_2}$, … $\mathbf{u_n}$ should be as independent as possible. For a noise-free ICA model, **W** should be the inverse of **A**, and **U** should be equal to **S**, according to the following equation:

$$\mathbf{U} = \mathbf{WX} = \mathbf{W}(\mathbf{AS}) = \mathbf{S} \qquad \textbf{(II-5)}$$

The mixing matrix **A** can then be calculated as:

$$\mathbf{A} = \mathbf{XS^T}(\mathbf{SS^T})^{-1} \qquad \textbf{(II-6)}$$

2.2.3. Multivariate curve resolution-Alternating least squares

Multivariate curve resolution-alternating least squares (MCR-ALS) is a well-known resolution method [81; 82] which has the objective of decomposing an original matrix **X** (*n* samples or rows and *p* variables or columns) of a multi-component system into the underlying bilinear model which assumes that the observed spectra are a linear combination of the spectra of the pure components in the system:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad \textbf{(II-7)}$$

where **C** is the matrix of concentration profiles, $\mathbf{S}^{\mathbf{T}}$ the matrix of pure responses (i.e. spectra) and **E** contains the experimental error. In resolution of spectroscopic images, **X** is the matrix of the unfolded image, **C** contains the concentration profiles that, conveniently refolded, show the distribution maps of each image constituent and $\mathbf{S}^{\mathbf{T}}$ contains the associated pure spectra [83]. In order to provide chemically meaningful profiles (i.e. pure spectra and distribution maps) and to strive for a unique MCR-ALS solution, several constraints must be properly chosen during the iterative calculation process (non-negativity, equality…) [84-86].

MCR-ALS must be initialised by a first estimate of **C** or $\mathbf{S}^{\mathbf{T}}$ matrix. Initial estimates can be manually filled where pure spectrum of each constituent is known but generally, a mathematical approach is applied. SIMPLISMA (Simple-to-use interactive self-modeling mixture analysis) [87], orthogonal projection approach (OPA) [88], independent component analysis (ICA) [89] or evolving factor analysis (EFA) [90] were used on spectroscopic data to identify pure signals in a mixture dataset.

During iterative process, figures of merit are the lack of fit (lof) and the explained variance (R^2). The lack of fit is used to check if the experimental data were well fitted by the MCR-ALS procedure. These two criteria are calculated as follow:

$$lof(\%) = 100\sqrt{\frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} X_{i,j}^2}} \qquad \textbf{(II-8)}$$

$$R^2 = \frac{\sum_{i,j} X_{i,j}^2 - \sum_{i,j} e_{i,j}^2}{\sum_{i,j} X_{i,j}^2} \qquad \textbf{(II-9)}$$

where $X_{i,j}$ is the input element of the original matrix **X** and $e_{i,j}$ the related residual element after using the MCR-ALS model.

3. Identification of a low dose compound

3.1. Definition of a low dose compound

The main objective of this thesis is the detection of a low dose compound in a pharmaceutical drug product by using Raman microscopy. In a large point of view, it can be generalised to the detection of a scarce sample in hyperspectral dataset and it can be extended to other applications (example: identification of a contaminant in food engineering). A scarce sample can be defined by a compound which has low spatial distribution and low spectral contribution in the data cube.

By definition, hyperspectral dataset are characterized by spectral and spatial dimensions. Regarding the spatial aspect, a specific compound can be distributed in most pixels of the image (the distribution of this compound can be considered as homogeneous) or in a few pixels of the image (the distribution of this compound can be considered as heterogeneous). Regarding the spectral aspect, the compound can provide high or low spectral contributions, depending on its concentrations or proportions in a spectrum, or depending on its absorptivity or spectral responses. As it is shown in Table II-1, four different cases can be found and observed for a specific compound in an image. Two of them (high spectral contribution in most pixels and high spectral contributions in a few pixels) can be easily tackled since the spectral information is highly present in several image pixels, i.e. several spectra. In this work, only the case of a compound which has low spatial and spectral contributions was studied.

		Spatial	
		In most pixels	**In a few pixels**
Spectral	**High spectral contributions**	Pure spectra and distribution maps can be easily calculated	Pure spectra and distribution maps can be easily calculated
	Low spectral contributions	Low spectral contribution of the compound	**Low spatial and spectral contributions of the compound**

Table II-1 Spatial and spectral contributions of a compound

In a pharmaceutical sample, a low dose compound can be an active (low concentrated drug substance, polymorph, impurity…) or an excipient (lubricant…).

3.2. The sampling aspect

In the pharmaceutical environment, the main objective of Raman chemical imaging is to study the distribution of actives and excipients in tablets or powders. Even if a pharmaceutical drug product is included within the quality specifications, the different compounds can be considered as non-homogeneously distributed in the tablet, leading to a possible sampling error if the entire image of the sample is not acquired. Because the whole tablet is not perfectly homogeneous, acquisition of different areas could provide various results [91].

Raman microscopy has been previously tested to study the identification of a low-content active pharmaceutical ingredient. In [92], tablets were prepared with two forms of API which one is considered undesirable and lower than 1% w/w. Authors focused on the number of image spectra to acquire in order to ensure the spectral detection of the low-concentrated form. The probability of observing at least one spectrum of a low dose compound can be calculated as follow:

$$x_{ld} \sim Bin(n, c) \qquad \textbf{(II-10)}$$

and

$$Prob(x_{ld} > 0) = 1 - (1 - c)^n \qquad \textbf{(II-11)}$$

Where "bin" stands for binomial distribution, x_{ld} is the number of spectra of the low dose compound found from n spectra, and c the concentration of the low dose spectra. For example, for a 0.5% w/w low dose compound in a formulation, there is a probability higher than 99% to find a spectrum if more than 1000 spectra are acquired (Figure II-7).

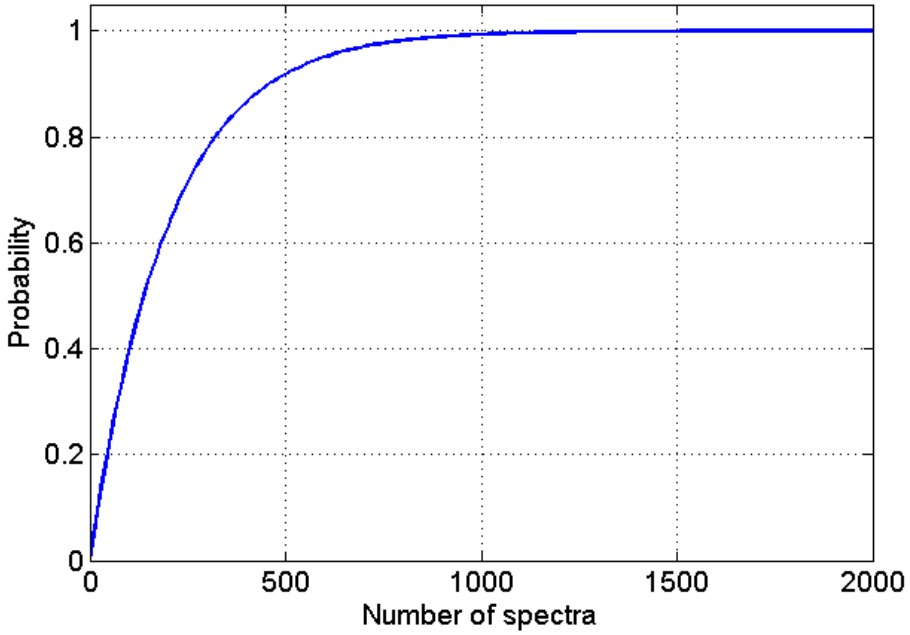


Figure II-7 – Probability of finding at least one spectrum of a 0.5% w/w low dose compound

Even if these studies have shown the critical aspect of the number of spectra in an image, the statistical approaches developed did not consider a critical parameter of an acquisition: the spatial resolution. With chemical imaging, Raman measurements can be performed at a macroscopic or a microscopic scale. When a macroscopic scale is used, the spatial resolution can be generally considered as lower than the particle size of each compound. Therefore, the measured signal may not be representative of a pure pixel composition, and can be a mixture of different compounds. This occurs because of the light penetrating deep into the sample. The acquired signal at a specific pixel position is not originated only from a small and confined volume on the surface of the tablet but also from under the surface and the sides, depending on the apparatus and the sample. In theory, the higher the magnification, the smaller the dissipation of the light and the sampling area which leads to a pure compound identification [93]. Because Raman signals from various compounds of the tablet normally interfere [92], spectral modifications can be difficult to identify in the case of a low dose compound.

The probability of finding spectra (or pixels) of a given constituent is related to its concentration in the formulation, to its distribution in the tablet, to its scattering coefficient, to apparatus and acquisition parameters. Even if optimization of the chemical imaging system (for example by using a high spatially resolved spectrometer) appeared as a straightforward solution to detect a low dose compound, it increases significantly the number of points and hence, the time required for image acquisition.

By considering the spectral variations, several compounds can have interferences, with overlapped Raman bands, which can make the identification of a constituent harder, especially in the case of a low dose compound. Therefore, multivariate data analysis of the spectral dataset can significantly improve identification and detection of a compound. But in practice, the precision and sensitivity of qualitative or quantitative analysis is very sensitive to both the spectrometer and sampling errors.

3.3. Data analysis aspect

For a lot of compounds, Raman spectroscopy provides spectrum with sharp and well-defined Raman bands. Generally, active responses are much stronger than those from excipients so that even low concentrations can be satisfactory detected. For those reasons, the distinction of active among the components of a formulation can be manually and visually performed by the analyst [94; 95] using univariate observation at a single variable (i.e. Raman shift) or by calculating surface ratio of Raman bands.

However, in most applications, Raman bands are overlapped and a direct interpretation of the spectra is not possible. Therefore, chemometric tools appeared as the only solution to extract useful information from the acquired signals. Several chemometric methods have been developed on vibrational dataset or hyperspectral imaging data cube (see Chapter II, paragraph 2) and most of them have studied the distributions of "sufficiently concentrated" pharmaceutical compounds, using prior knowledge on the studied formulation.

In some cases, the detection of a low dose compound can be useful to ensure the product quality or to improve the development of a product. In the literature, it has been previously studied from usual spectroscopic data by using bulk measurements [96; 97] and chemometric tool such as the PLS regression [98-100] and some of them focused on the detection limit of the analytical method [101]. The use of the net analyte signal (NAS) [102; 103] pre-processing appeared as an interesting tool to accurately resolve the analyte signal of a low dose compound and allow the construction of a quantitative model [104]. Several adaptations of these approaches can be considered, depending on the spectral basis (i.e. space containing the contributions to be removed) used for projecting the original dataset.

With the definition provided in section 3.1, a low dose compound can be viewed as a product with low spectral variance (i.e. low spectral contribution) within the entire dataset. The variance is one of the moments of a distribution. In theory, it describes how far a set of samples is spread out around the mean. In this work, because the data are not centered, the variance can be

associated with the dispersion of samples around a predefined value. Due to low spectral contributions of the low dose compound, and because it is only present in a few spectra (i.e. pixels), a visual identification of its distribution is not possible and usual chemometric method, mainly based on statistical moment decompositions, may encounter some difficulties to extract the associated information. Therefore, the detection of a low dose compound by using Raman microscopy and chemometrics appeared as a real challenge and, to our knowledge, it has not been studied in a previous work.

3.4. Contributions of the thesis

The present work uses Raman microscopy to study the distribution of actives and excipients in a pharmaceutical drug product. It focuses on the application of chemometric tools to identify both major and minor compounds of a pharmaceutical formulation, including spectral features and distribution maps of each product. The detection of a low dose compound in a tablet is the common thread of this thesis. It was defined above as a product with low spatial and spectral contributions, meaning that the information is contained only in few pixels of the image and mixed with the other compound spectra or scattered in noise contribution.

With usual chemometrics methods, hyperspectral image analysis can be viewed as the resolution of the following equation: $\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathbf{T}}$ where $\mathbf{X}$ is the initial dataset, $\mathbf{C}$ the matrix of concentrations and $\mathbf{S}$ the matrix of pure spectra. $\mathbf{C}$ and $\mathbf{S}$ can be calculated without prior knowledge by using blind source separation methods or with prior knowledge by using resolution methods. In this work, ICA and MCR-ALS, which have been previously applied on spectroscopic measurements and hyperspectral imaging to provide spectral features and distribution maps, were used. In both cases, the decomposition of statistical moments (variances or cumulants) was required. Considering the studied case of the low dose compound, we can make the hypothesis that, because these algorithms are mainly based on the decomposition of statistical moments, identification of this product within hyperspectral dataset can be difficult and different improvements or adjustments should be required.

In chapter III and chapter IV, the hypothesis presented above is challenged by applying ICA and MCR-ALS on hyperspectral image of a pharmaceutical tablet to provide the distributions of actives and excipients. The studied sample includes a lubricant which corresponds to the low dose compound. By applying ICA and MCR-ALS as usual, without any modifications of the calculation process, it can be assumed that these algorithms are not able to extract the low dose compound contributions. Some improvements and modifications of these two algorithms are

proposed and tested. In chapter III, the use of over-segmented ICA model is described. In chapter IV, modifications of the filtering process prior than the iterative MCR-ALS process are tested. Chapter III and chapter IV validate the difficulty of extracting the low dose compound contribution by using algorithms based on the decomposition of statistical moments. Rather than using the statistical moments, this thesis investigates alternatives method based on the signal space. It describes the P-dimensional space (one axis per variable) in which the observations can be represented as vectors. It ensures the detection of a compound without requiring important variations between samples (or pixels) and it appears as particularly suitable for the studied case of a low dose compound.

Therefore, second part of the thesis, constituted of chapter V and chapter VI, uses orthogonal projections to improve the performance of MCR-ALS algorithm by calculating a constraint based on signals, and provides a new approach for the detection of a low dose compound in a pharmaceutical drug product.

In chapter V, the work focuses on MCR-ALS calculation and especially on the optimisation of the spatial constraint frequently used to improve the resolution. Indeed, MCR-ALS requires the use of constraint to reduce intensity or rotational ambiguities and to tend to a unique solution. Equality constraint, based on local rank information, was previously studied. However, as the usual method applies singular value decomposition on several spectra, it requires a sufficient level of differences between samples. Since it is based on the use of second central moment (i.e. variance decompositions), limitations of this approach are reached in the case of a low dose compound. A new methodology to set up absence/presence maps is proposed. It is based on orthogonal projection to a basis containing all the spectral variability other than the one of the compound of interest. It can only be applied in situations where the space of interferences can be well-defined, and thus, it requires to know the sample composition beforehand.

In the previous chapters, the pharmaceutical composition is supposed to be known by the analysts. However, in some applications, drug products contained in tablets or powders are not known. In chapter VI, an iterative method for compound detection in an unknown drug product is proposed. The proposed methodology requires a spectral library, spectral distances and orthogonal projections to iteratively detect the compound of a mixture matrix. Again, this iterative method is only based on the spectral space, without requiring information between samples (or spectra). The approach is tested and discussed on a pharmaceutical drug product including a low dose compound but conclusions and proposed approaches can be extended to other similar applications.

Chapter III: Use of blind source separation approach for pure spectra determination and spatial distribution of constituents

1. Introduction 28

2. Materials and methods 30

2.1. Samples 30

2.2. Raman imaging system 30

2.3. Pre-processing 30

2.4. Independent Component Analysis (ICA) 31

2.5. Data analysis 32

3. Results & discussion 32

3.1. Selection of number of independent components 32

3.2. Distribution of API 33

4. Conclusions 42

Preamble

In this chapter, a Raman hyperspectral image of a commercialised tablet is studied. The objective is to examine the distribution of active principal ingredients and excipients within the tablet. A lot of chemometrics tools have been previously studied to extract distribution maps and most of them require prior knowledge for calculation or data interpretation. In this work, we want to focus on the use of independent component analysis (ICA), a blind source separation method, to extract interpretable pure signals (Figure III-1).

With ICA, each row of a data matrix is considered to be a sum of pure source signals, neither the source signals, nor their proportions being known. ICA aims to extract these pure sources, underlying the observed signals, by maximization of their non-Gaussianity, as well as their concentration in each mixture. As this approach can be used without pure spectra knowledge, this is of a huge interest comparing with other chemometric algorithms.

Since ICA results depend on the number of independent components used in the model, this criterion is considered as a critical parameter. Most of the time, it is determined based on prior knowledge concerning the studied case. In order to avoid this manual selection, an innovative method using the comparison of signals between spectral blocs is used. Being a critical parameter of the ICA model, the number of ICs is intentionally modified, simulating under-decomposition or over-decomposition, in order to test the effect on results.

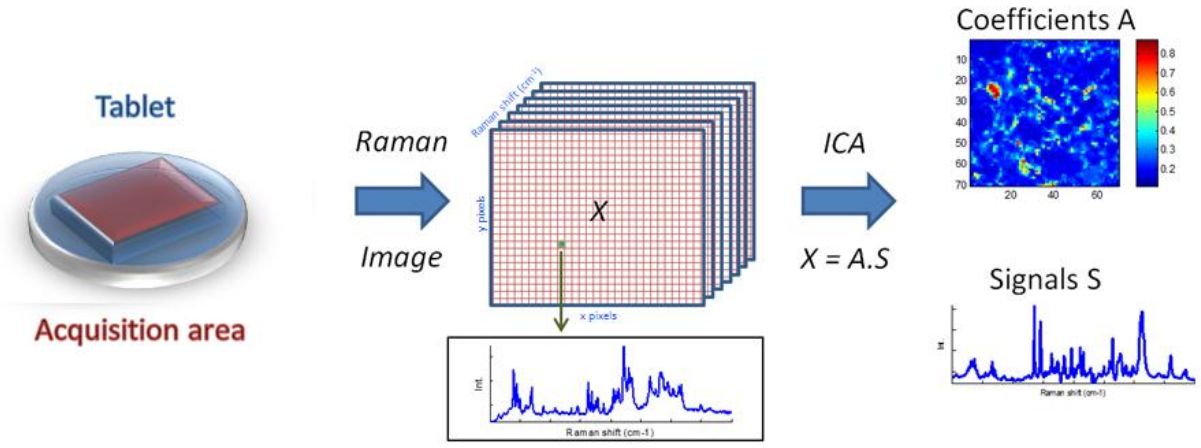


Figure III-1 Graphical representation of the tested approach

In the case of a low dose compound, pure signal detection using ICA seems to be a real challenge. Indeed, due to its low spectral contribution and to its presence in a few pixels, the variability linked to this product is weak. The method is tested and discussed on a tablet manufactured with a low dose lubricant.

This chapter is the reproduction of **Art. I** published in the Journal of Pharmaceutical and Biomedical Analysis in 2015.

APPLICATIONS OF INDEPENDENT COMPONENT ANALYSIS ON RAMAN IMAGES OF A PHARMACEUTICAL DRUG PRODUCT: PURE SPECTRA DETERMINATION AND SPATIAL DISTRIBUTION OF CONSTITUENTS[1]

1. Introduction

In recent years, chemical imaging has become an emerging technique that integrates conventional imaging and spectroscopy to combine spatial and spectral information from a sample [15]. The use of vibrational spectroscopies such as near infrared or Raman is particularly appreciated within the pharmaceutical research and development environment. Indeed, vibrational spectroscopy technologies on solid pharmaceutical samples have many advantages such as the rapidity of analysis, the non-destruction of the sample and the possibility to perform an analysis without using solvents. The spatial information provides useful information on product processing, for formulation development or to control the quality of an existing drug product. Indeed, the distribution of actives or excipients within a specific formulation becomes an important quality control parameter.

Several applications of Raman spectroscopy have been published and the potential of this technique is widely accepted [105]. The use of Raman spectroscopy for the detection of trace cristallinity [106] and the determination of active content within pharmaceutical capsules [107], are of great interest for the development and the quality control of a formulation. Moreover, hyperspectral imaging shows considerable promise for providing information in diverse fields such as remote sensing [108] for interpretation of experimental spectroscopic images from the geographical region of Cuprite, foods and agriculture [109] for analysis of cucumber leaves and pharmaceuticals for analysis of solid dosage forms [110] or the detection of polymorphic forms in tablets [93].

Coupling spectroscopy and imaging generates a huge amount of data. Most of the time, the image cube is unfolded into a data matrix and to extract the maximum of information, it is necessary to use multivariate data analysis methods and spectral decomposition techniques [111]. Standard chemometric tools such as principal component analysis [112], cluster analysis [113], classical

[1] Mathieu Boiret, Douglas N. Rutledge, Nathalie Gorretta, Yves-Michel Ginot, Jean-Michel Roger. **Applications of independent component analysis on Raman images of a pharmaceutical drug product: pure spectra determination and spatial distribution of constituents.** *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 90 (2014) 78-84.

least squares [48] and multivariate curve resolution [114] have previously been described in the literature on Raman datasets.

Independent component analysis (ICA) is a blind source separation algorithm [79] particularly appreciated for the decomposition of spectroscopic data. Its ability for spectral decomposition of UV-VIS spectra has already been evaluated [115]. Wang et al. [116] also highlighted that ICA can be used as a blind source separation technique to extract pure component information from various measured analytical signals such as mass spectra, mid-Infrared spectroscopy spectra or chromatograms. In this article, ICA was applied on a promising technique for pharmaceutical drug product analysis: the Raman spectroscopy. In ICA, each row of the data matrix is considered to be a sum of pure source signals, neither the source signals, nor their proportions being known. ICA aims to extract these pure sources, underlying the observed signals, as well as their concentration in each mixture. Source signals are assumed to have a definite structure, and so their intensity does not have a Gaussian distribution. On the other hand, although the distributions of independent signals are not Gaussian, their sum tends towards a Gaussian distribution. ICA aims to extract the pure source signals by maximization of their non-Gaussianity [117].

In this paper, a commercial pharmaceutical tablet was analysed by Raman chemical imaging. The objective was to extract interpretable pure signals using ICA, in order to examine the distribution of active principal ingredients (API) and major excipients. ICA approach can be used without pure spectra knowledge. The direct data analysis of the image is a huge advantage comparing with the usual Chemometric algorithms. This approach can become a useful tool for quality control of a pharmaceutical drug product or to analyse a product with an unknown composition. As a method based on decomposition of the original data matrix, the number of independent components is a critical step of this algorithm. Usually the number of independent components to extract is determined based on prior knowledge concerning the formulation [118]. In order to select the best number of independent components, innovative tools previously developed and published were used in this study. Each calculated source signal was compared with the pure spectra of the constituents and the distribution of the compound in the tablet determined. Being a critical parameter of the ICA model, the number of ICs was intentionally modified, simulating under-decomposition or over-decomposition, in order to test the effect on results.

2. Materials and methods

2.1. Samples

A commercial coated tablet of Bipreterax® was used for the study. Bipreterax® is used for arterial hypertension treatment and is commercialised by "Les Laboratoires Servier". It is also known as Perindopril (active principal ingredient 2 or API 2) / Indapamide (active principal ingredient 1 or API 1) association and contains respectively 4mg of API 2 and 1.25mg of API 1 in the commercial drugs. Actives are known to have several solid state forms, but only one of them is present in this formulation. Major core excipients are lactose monohydrate, microcrystalline cellulose (Avicel) and magnesium stearate. In order to analyse the tablet core, the coating was removed by eroding the sample with a Leica EM Rapid system (Leica, Wetzlar, Germany). A visual examination of the tablet did not provide any information concerning the distribution of the different compounds within the tablet.

2.2. Raman imaging system

The image was collected using a RM300 PerkinElmer system (Perkin Elmer, Waltham, MA) and the Spectrum Image version 6.1 software. The microscope was coupled to the spectrometer and spectra were acquired through it with a spatial resolution of 10µm in a Raman diffuse reflection mode. Wavenumber range was 3200–100 cm^{-1} with a resolution of 2 cm^{-1}. Spectra were acquired at a single point on the sample, then the sample was moved and another spectrum was taken. This process was repeated until spectra of points covering the region of interest were obtained.

A 785nm laser with a power of 400mW was used. Two scans of two seconds were accumulated for each spectrum. An image of 70 pixels per 70 pixels corresponding to 4900 spectra was acquired for a surface of 700µm by 700µm.

2.3. Pre-processing

Data were pre-processed in order to remove non-chemical biases from the spectra (scattering effect due to non-homogeneity of the surface, interference from external light source, spikes due to cosmic rays, random noise). First of all, data were spike-corrected in order to reduce the effect of cosmic rays [61]. Next, the spectral range was reduced in order to focus only on the region of interest, corresponding to a Raman shift from 1800cm^{-1} to 200 cm^{-1}. Reduced spectra

were pre-processed by standard normal variates correction (SNV) [72] in order to reduce the effect of baseline variations and uninformative variations in global spectral intensity.

2.4. Independent Component Analysis (ICA)

ICA is one of the most powerful techniques in blind source separation [119]. It has been developed to extract the pure underlying signals from a set of mixed signals in unknown proportions. Considering a noise-free ICA model, a matrix **X** (*n* x *m*) of *n* spectra and *m* variables (Raman shift) is decomposed as a linear generative model by the following expression:

$$\mathbf{X} = \mathbf{AS} \qquad \textbf{(III-1)}$$

Where **S** is a (*k* x *m*) matrix of *k* independent source signals called the independent components and **A** is a (*n* x *k*) mixing matrix of coefficients or proportions of the pure signals in each mixed signal of **X**.

The objective of ICA is to estimate a set of vectors that are as independent as possible, and the mixed signals in **X** can then be expressed as linear combinations of these independent components (ICs). It attempts to recover the original signals by estimating a linear transformation, using a criterion which reflects the statistical independence among the sources.

To solve the previous equation (Eq. III-1), an unmixing matrix **W** based on the observation of **X** needs to be calculated. The output **U**, constituted by the independent component $\mathbf{u_1}$,$\mathbf{u_2}$, … $\mathbf{u_n}$ should be as independent as possible. For a noise-free ICA model, **W** should be the inverse of **A**, and **U** should be equal to **S**, according to the following equation:

$$\mathbf{U} = \mathbf{WX} = \mathbf{W}(\mathbf{AS}) = \mathbf{S} \qquad \textbf{(III-2)}$$

The mixing matrix **A** can then be calculated as:

$$\mathbf{A} = \mathbf{X}\mathbf{S}^{\mathbf{T}}(\mathbf{S}\mathbf{S}^{\mathbf{T}})^{-\mathbf{1}} \qquad \textbf{(III-3)}$$

Lots of algorithms are available to perform ICA calculations such as FastICA [120] or Radical [121]. In this paper, the JADE (Joint Approximate Diagonalization of Eigenmatrices) algorithm was used [122]. Compared with other methods based on parameter optimization, the JADE algorithm performs matrix diagonalizations, and therefore does not involve an optimization procedure [123].

The ICA_by_blocks algorithm [124] was used to determine the optimal number of signals to extract. This method starts by splitting the initial data matrix **X** into *B* blocks of samples (with approximately equal numbers of rows). Note that the samples in each block have to be representative of the whole dataset. ICA models are then computed with an increasing number of ICs for each block. To ensure the same signs of the ICs of the different models, the signs of the vector **A** (and therefore the corresponding **S**) are adjusted so that the most intense value in each vector of **A** is positive. ICs corresponding to true source signals should be found in all representative subsets of samples, or row blocks, of the full data matrix. These ICs should be strongly correlated.

2.5. Data analysis

Data analysis was performed by using Matlab R2012a software. The Matlab code of the JADE algorithm was downloaded from the web site in ref. [125].

3. Results & discussion

3.1. Selection of number of independent components

Determination of the number of ICs for ICA decomposition is a critical step of the data analysis. Indeed, calculating too few ICs results in non-pure signals, whereas calculating too many ICs can decompose pure signals into several contributions. The ICA_by_blocks method was applied by splitting the dataset row-wise into two blocks and by performing ICAs on each block. Sample selection to create the two subsets was done by using a "venetian blind" procedure. Each test set is determined by selecting every b^{th} (number of blocks) object in the dataset, starting at object number one. ICA models were calculated for both blocks with from 1 IC to 20 ICs. ICs were compared in each block by calculating the correlation coefficients between all pairs of signals from both blocks for a given model. The highest-dimensional model for which ICs obtained in a block were similar to ICs obtained in another block indicates the optimal number of ICs to extract from the data under study. Figure III-2 shows that the lowest correlation between signals significantly decreases after 9 ICs, which was therefore considered as the optimal number of component for the decomposition of the dataset. The initial drop after 4 ICs and then after 7 ICs is assumed to be due to the fact that the ICs are not extracted from the two data blocks in exactly the same order.

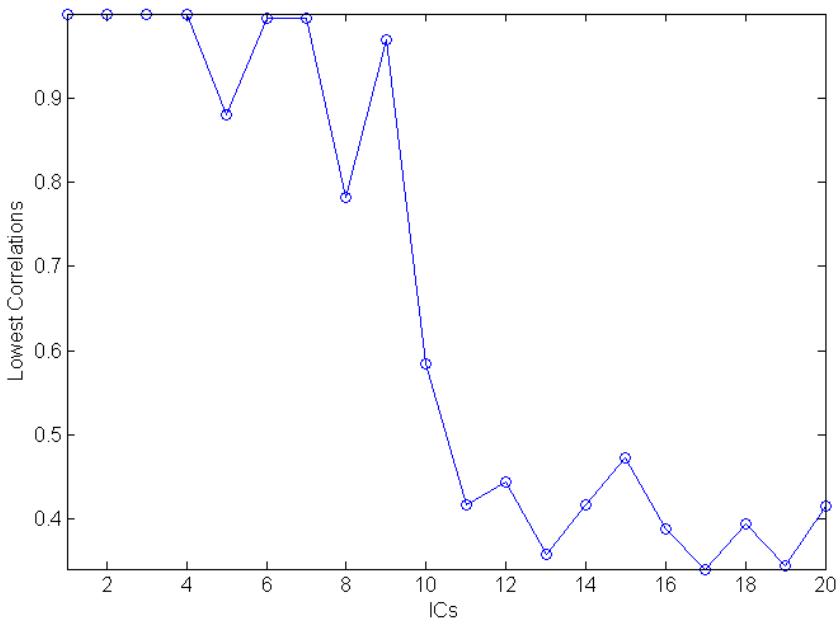


Figure III-2 Lowest correlation between signals obtained using ICA_by_blocks. The lowest correlation obtained using the ICA_by_blocks approach significantly decreases after 9 ICs, which was considered as the optimal number of component for the decomposition of the dataset

Since the sample contains five compounds and supposing that the five spectra are independent and that the acquired mixture spectra are linear combinations of the pure spectra, five ICs should have been sufficient. In this example, in contrast with the theoretical decomposition, four more components were used to build ICA models. Physical effects such as particle size variation or fluorescence of a compound could explain this "over-decomposition" of the dataset.

3.2. Distribution of API

An ICA model based on the JADE decomposition with 9 ICs was calculated on the unfolded, SNV pre-processed data cube. The matrices of the proportions, **A**, for each signal, **S**, were then folded back in order to obtain a representation of the spatial distribution of each independent component. In Figure III-3, different textures of images can be observed. Indeed, IC1, IC6 and IC9 show very specific inhomogeneous distributions with agglomerates. Considering the different scales of score images, IC2, IC3, IC4, and IC5 have similar textures (or distributions) such as IC7 and IC8 which are the same as that in IC1. It can also be seen that the distributions observed in these two sets of images are complementary, indicating that these two sets of Independent components occupy complementary regions in the tablet. In order to associate an independent component with a chemical compound, the calculated signals were examined.

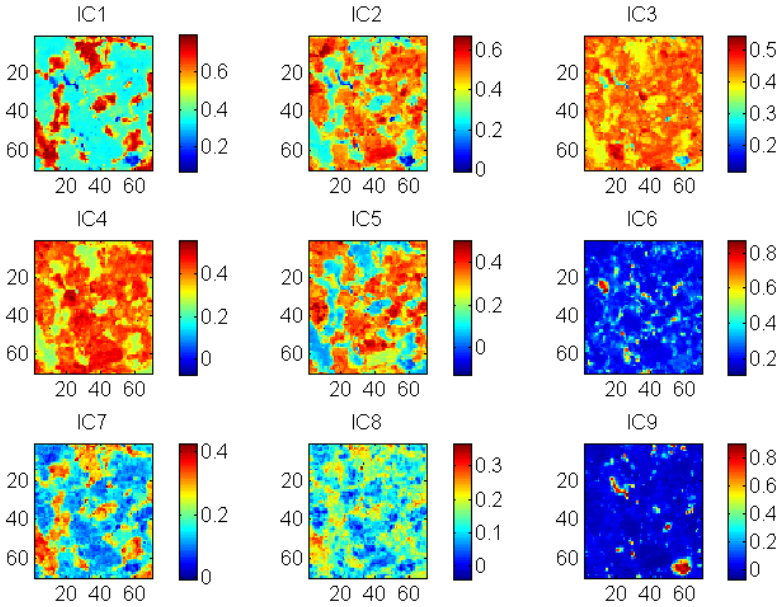


Figure III-3 Proportions coefficients (A) of each IC. Images correspond to the proportions coefficients (A) of a 9 ICs model. A red color corresponds to a high value whereas a blue color corresponds to a low value.

Figure III-4 shows the 9 signals calculated by ICA. Signals from IC1, IC2, IC3, IC4, IC5, IC6 and IC9 look like well-defined Raman spectra with no baseline shift due to fluorescence effects whereas the signals in IC7 or IC8 contain noise and baseline variations which could be explained by a fluorescence effect. In theory, and supposing the independence of each spectrum within the formulation, 5 ICs should have been sufficient for the matrix decomposition. However, 9 ICs were determined to be present, possibly due to physical effects, or interactions between constituents. Considering the simplicity of the preprocessing method applied on the Raman spectra (spike correction, selection of a specific range and SNV), the quality of the calculated signals was sufficient and perfectly suitable for analytical interpretation.

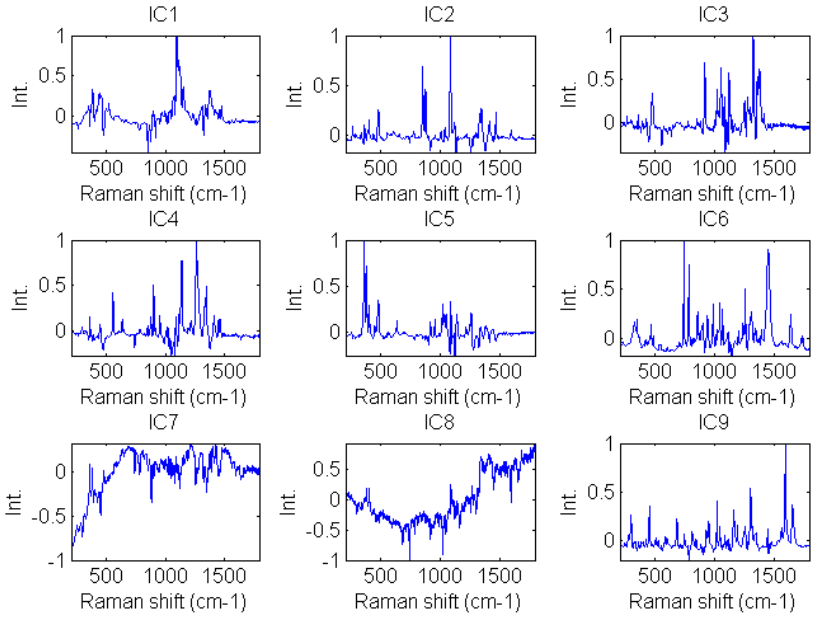


Figure III-4 Signals, S, of the ICA model. These signals correspond to the calculated signals (S) of a 9 ICs model.

The spectra for the known constituents in the tablets are plotted in Figure III-5. Even though the spectra of all compounds are very different, lots of Raman bands are overlapped. A mixture spectrum is a combination of these spectra, given the presence of each constituent in any specific pixel of the image. In order to interpret the ICA results, the correlation coefficients between the ICA signals and the pre-processed spectra of the compounds were calculated. Results can be found in Table III-1.

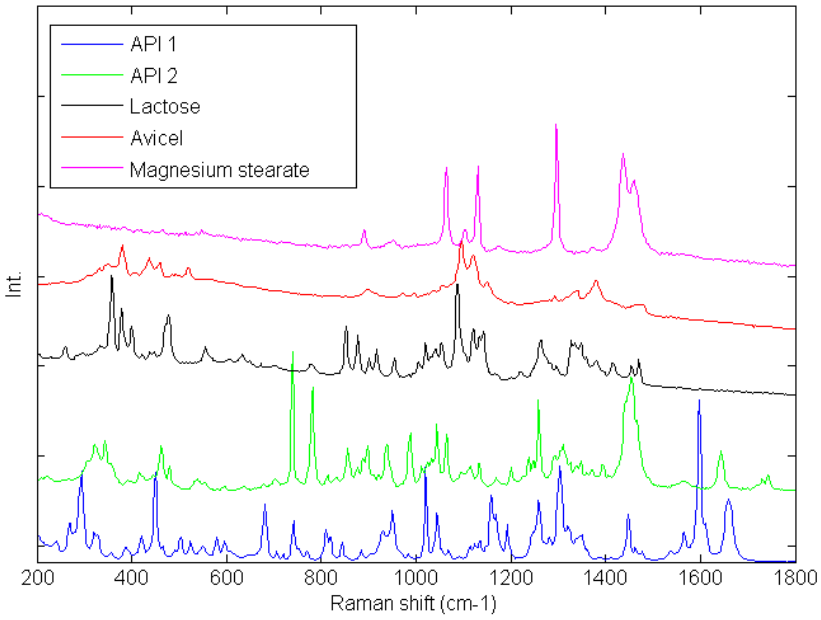


Figure III-5 Pure spectra of the drug product constituents. In blue API 1, in green API 2, in black lactose, in red avicel and in magenta the magnesium stearate. Relative intensities were used as the spectra were split for a better observation.

The comparison between the calculated signals and the true spectrum of each compound shows that only two ICs are directly linked to the drug product constituents. For each component, the highest correlation was highlighted with bold characters in Table III-1.

Pure spectrum	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	IC9
API1	0.01	-0.09	0.07	0.14	-0.04	0.13	0.21	0.18	**0.92**
API2	0.06	-0.01	0.11	0.08	0.03	**0.96**	0.08	0.10	-0.06
Lactose	0.25	**0.44**	**0.23**	**0.25**	**0.47**	0.00	0.36	0.45	-0.17
Avicel	**0.49**	0.15	0.06	0.02	0.20	-0.07	**0.38**	**0.61**	-0.20
Magnesium Stearate	0.20	0.00	0.01	0.04	0.04	0.41	0.32	0.23	-0.12

Table III-1 Correlation coefficients between the ICA signals and the pre-processed true compound spectra. The comparison between the calculated signals and the true spectrum of each compound shows that only two ICs are directly linked to the drug product constituents. For each component, the highest correlation was highlighted with bold characters.

No high correlations were found for Magnesium stearate. Two very high correlations were highlighted between the pure spectra and the calculated signals (respectively 0.92 between IC9 and the active principal ingredient 1 and 0.96 between IC6 and the active principal ingredient 2). As is shown in Figure III-6 and Figure III-7, the calculated signals (IC9 and IC6) are in effect very similar to the pure spectra of API 1 and API 2. The refolded images of the corresponding proportions, **A**, therefore reflect the distribution of these two compounds. As can be seen in Figure III-3, the distribution of active principal ingredients is not perfectly homogeneous and agglomerates are observed.

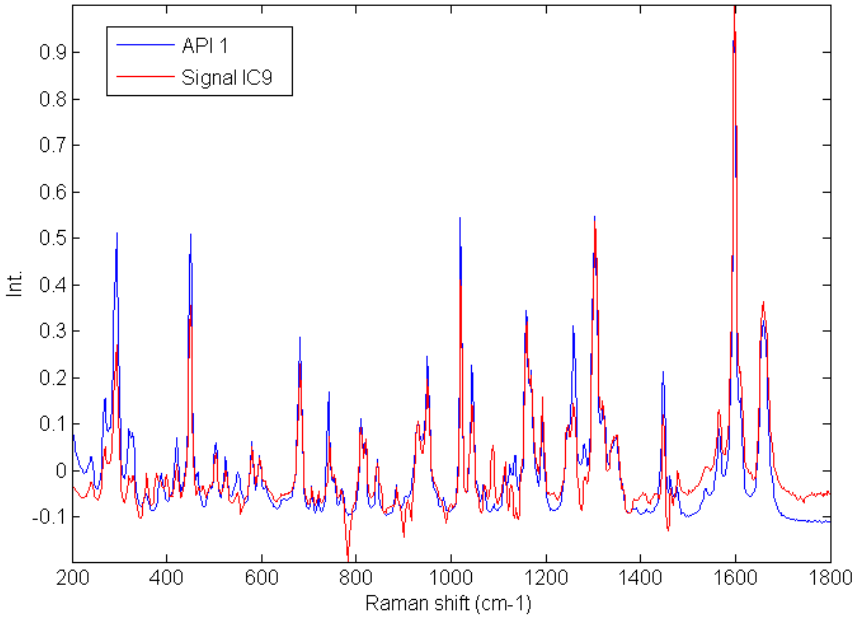


Figure III-6 Calculated signal of independent component 9 superposed on the spectrum of API 1. Comparison between API 1 spectrum and IC9 signal. The correlation between the two signals is equal to 0.92.

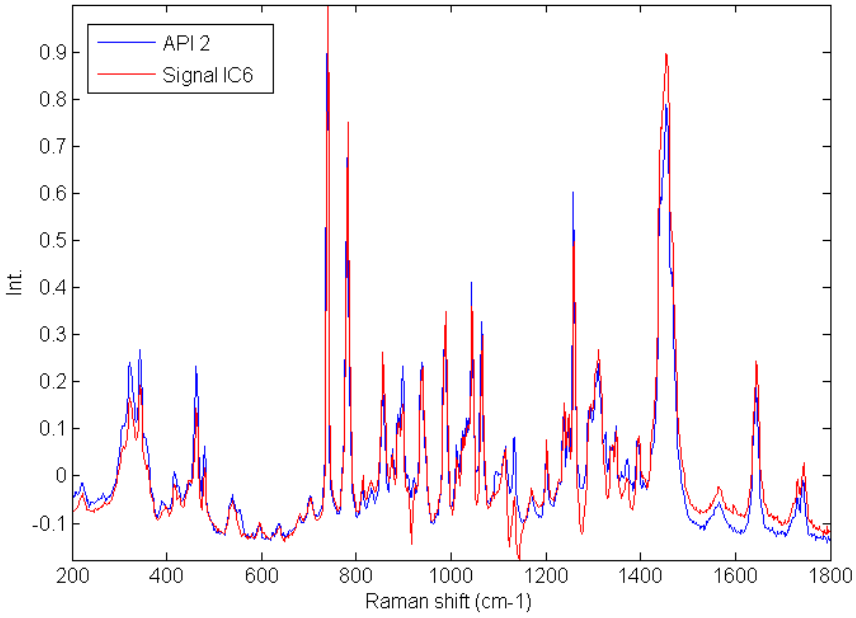


Figure III-7 Calculated signal of independent component 6 superposed on the spectrum of API 2. Comparison between API 2 spectrum and IC6 signal. The correlation between the two signals is equal to 0.96.

IC1 is mainly correlated with the spectrum of Avicel. Specific bands due to the chemical bond vibrations are observed in this component (especially between 1250cm^{-1} and 1000cm^{-1}, spectral range linked to CC ring bond stretches and CO stretches). IC7 and IC8 are mainly correlated with the spectrum of avicel (0.38 for IC7 and 0.61 for IC8) but the correlation with lactose (0.36 for IC7 and 0.45 for IC8), magnesium stearate (0.32 for IC7 and 0.23 for IC8) and API1 (0.21 for IC7 and 0.18 for IC8) cannot be considered as non-significant. IC7 and IC8 signals are not well defined Raman spectra and contain principally noise or baseline variations which can explain these high correlations with several different products. As can be seen in Figure III-3, IC7 and IC8 have similar spatial distributions which are the same as that in IC1. Avicel is a microcrystalline cellulose powder which is known as a product providing a fluorescence effect with Raman, which could explain the contribution of IC7 and IC8.

As is shown in Figure III-8, IC2, IC3, IC4 and IC5 are linked to the lactose spectrum. Lots of lactose Raman bands are identified in these IC signals (for example band at 460cm^{-1} in signals 2, 3 and 5 due to various CCO and OCO bending modes, or band at 1088 cm^{-1} linked to the stretching vibration of the COC bridge). These 4 components gave their highest correlations with the lactose spectrum. However, these correlations were low (from 0.23 to 0.47) reflecting the decomposition of the pure spectrum into 4 components. The signal decomposition was

particularly significant in the low Raman shift spectral range. In this spectral region, coupled CC and CO vibrations rather than single functional group are mainly observed. By observing the refolded image of coefficients, note that the distribution of this product was very similar (considering the different image scales).

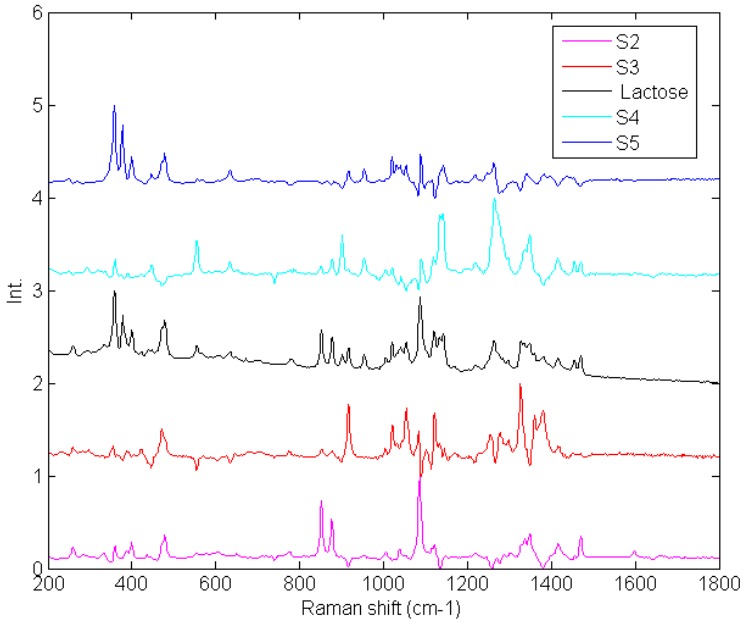


Figure III-8 Calculated signal of independent component 2, 3, 4, 5 plotted with the spectrum of Lactose. Comparison between lactose spectrum and IC2, IC3, IC4 and IC5 signals. The correlations between the signals are respectively equal to 0.44, 0.23, 0.25 and 0.47. The pure spectrum of lactose and the four calculated independent components are displayed. The pure spectrum was decomposed into four components.

The observed decomposition of the lactose information into separate Independent Components could be due to two phenomena. The first one is the physical effect. Indeed, lactose is known to have important particle size variations which can modify the light scattering and as a consequence the Raman spectra. These slight modifications could behave as independent phenomena and thus result in separate ICs. Moreover, the different combinations of vibrations could be interpreted by ICA decomposition as an independent variation. The second hypothesis is linked to the ICA decomposition itself. Indeed, as the formulation contains 5 compounds, the model may have mathematically over-decomposed the dataset by using 9 ICs.

In order to explore the ability of ICA to extract a pure signal from lactose, an ICA model was calculated with 5 ICs, which was the known number of constituents used to manufacture the

tablet. By comparing the 5 ICs with the pure spectra, a high correlation was found with lactose (R = 0.90), one with API1 (R = 0.95), one with API2 (R = 0.94), while a weak correlation was found with avicel (R = 0.39) and one signal contained noise and mixed pure contributions. The lactose contribution was therefore not divided among several components, as was observed when using 9 ICs. As is shown in Figure III-9, the calculated signal IC3 was very similar to the pure lactose spectrum. With 5 ICs, the decomposition of the original matrix was mainly due to chemical variations whereas the decomposition using 9 ICs included physical effects.

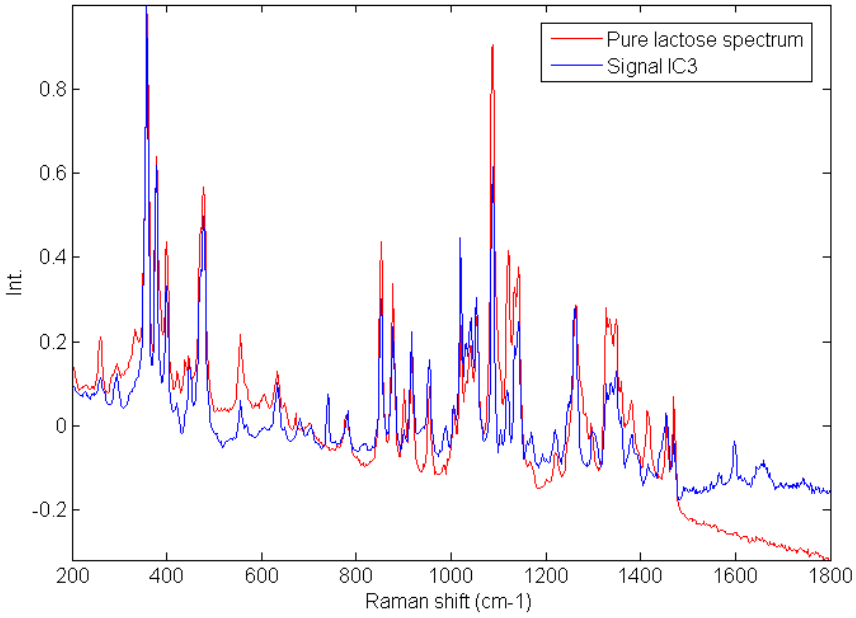


Figure III-9 IC3 signal from a 5 components ICA model superposed on the spectrum of lactose. Comparison between lactose spectrum and IC3 signal from a 5 components ICA model. The correlation between the two signals is equal to 0.90.

By observing ICA coefficients and signals, it can be seen that no information from the magnesium stearate was observed. The non-detection of this compound, frequently used as a lubricant in a pharmaceutical formulation, could be mainly due to its low concentration in the tablet (0.5 w/w%). Indeed, several hypotheses can be advanced to explain this lack of detection: the physical formulation of the product, the sensitivity of the spectroscopy or the failure of the ICA algorithm. As the analysed area does not represent the whole surface of the tablet and because of its low content, it is possible that the acquired spectra did not contain any magnesium stearate information. Moreover, the Raman contribution of the magnesium stearate

could be hidden by the contribution of the other constituents. In order to test the ability of ICA to detect and extract the information related to magnesium stearate, new models with more components and other pre-processing methods were tested (details not shown). By using a Savitzky-Golay pre-processing [59] and a model with 15 ICs, one signal (Figure III-10) was highly correlated (r = 0.87) with the pure spectrum of magnesium stearate and the distribution of the product can then be studied (Figure III-11). However, the quality of other signals significantly decreased. Pure spectra were divided among several components and the analytical meaning of each signal was not intuitive.

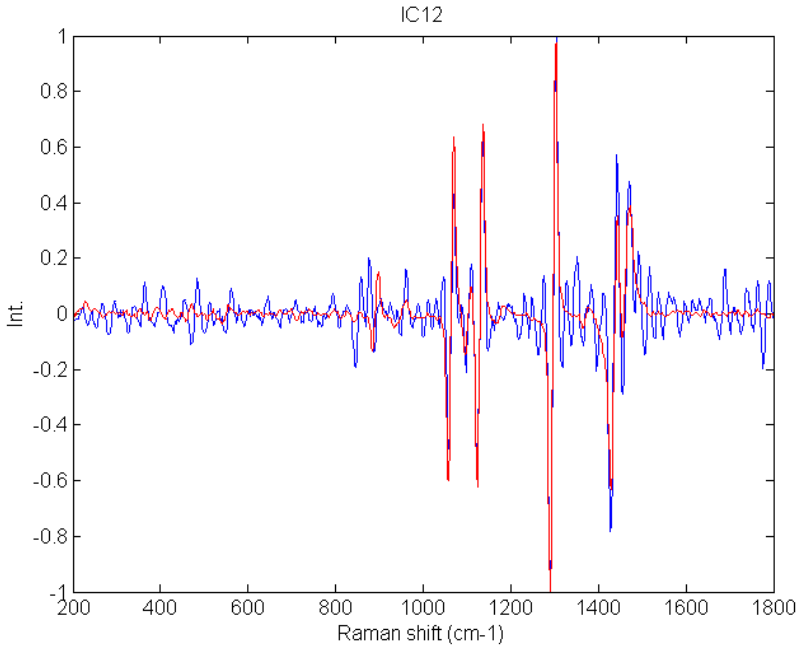


Figure III-10 IC12 superposed on the magnesium stearate spectrum from a 15 component ICA model. Comparison between magnesium stearate spectrum and IC12 signal from a 15 components ICA model. The correlation between the two signals is equal to 0.87.

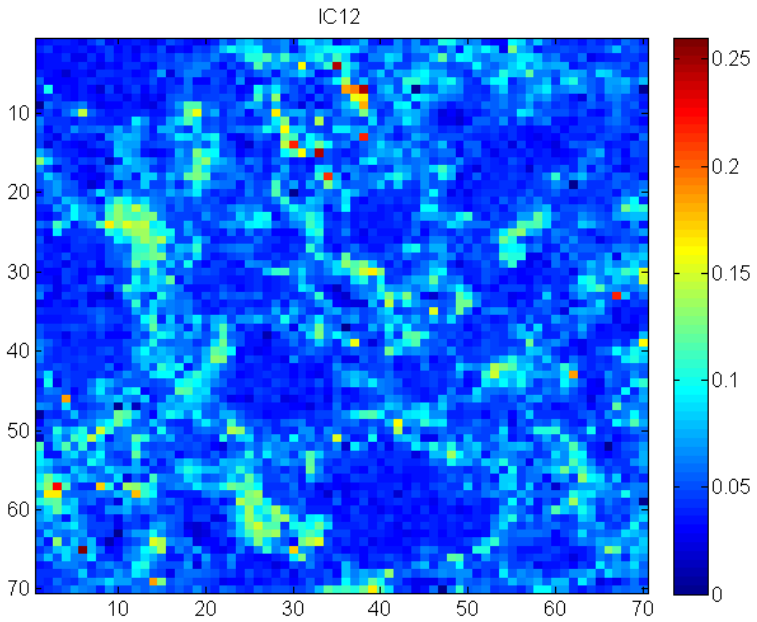


Figure III-11 Distribution of IC12 (magnesium stearate) from a 15 component ICA model. This component is highly correlated to magnesium stearate.

4. Conclusions

ICA was successfully applied on a Raman image of a commercial tablet. A representative image of the tablet was acquired and the spectrum of each pixel, which can be associated to a mixture of the different pure compounds, was pre-processed and analysed using the JADE algorithm to calculate signals and proportions with a specified number of components. This parameter was estimated by using the ICA_by_blocks method. This technique shows very good results to choose the most appropriate number of ICs on a real Raman dataset. It avoids arbitrary selection of this critical criterion.

This method gave good results to provide pure spectra of the active substances. Contribution of avicel was spread among 3 ICs. The first one was very similar to the pure spectrum of avicel whereas the two others were mainly fluorescence signals. Being a microcrystalline cellulose, avicel is known to be prone to fluorescence effects. The contribution of lactose was shared over 4 ICs which may be due to an over-decomposition of the original dataset or to physical contributions. In order to improve the pure lactose signal quality and based on knowledge of the product formulation, an ICA model was calculated using fewer ICs. The lactose contribution was then no longer divided among several signals but, the physical effects were no longer observed.

This should be contrasted with the fact that using an insufficient number of ICs leads to the non-detection of a low content compound, magnesium stearate. It has been shown here that using a very large number of components and another pre-processing method resulted in a well-defined ICA signal linked to magnesium stearate. It was then possible to examine the distribution of this low content product within the tablet. However, due to the over-decomposition of the dataset, other pure signals were divided among several components, which made the identification of each contribution within the tablet more difficult.

The ICA_by_blocks method was therefore a compromise between under- and over-decomposition. Even if the contribution of lactose or avicel were divided among several components, the spatial information obtained could be very useful for formulation development or to improve the quality control of pharmaceutical samples. New approaches, based on data fusion from ICA calculations to gather information from the same constituent, are under development and will be detailed in a future work.

Contributions of chapter III

In this chapter, ICA was tested in order to extract chemical pure signals from a supposed unknown pharmaceutical drug product. This blind source separation algorithm appeared as a powerful tool to extract pure signals without prior knowledge on the spectral dataset, i.e. the pharmaceutical formulation. By using these calculated signals, distribution maps of actives and excipients can be provided.

The impact of the critical parameter of ICA model, i.e. the selection of the number of components, was studied. It was shown that a model must be built by using an appropriate number of independent components. On the one hand, an under-fitted model was not able to extract all the pure signals of the studied formulation and the calculated signals were a mixture of several compounds. On the other hand, an over-fitted model extracted a number of components higher than the real number of products in the formulation. It provided signals divided among several components which were difficult to interpret. The ICA_by_blocks approach was a powerful alternative method to estimate the number of independent components for the decomposition.

Due to chemical or physical variability of each pure compound, the number of independent components which has to be calculated is often higher than the number of pure compounds in the formulation. For instance, lactose, a common excipient in the pharmaceutical development, is known to have important physical variability due to particle size variations. This variability can modify the light scattering and as a consequence the Raman spectra. Due to this spectral variation, the ICA model can mathematically over-decompose the dataset by providing a number of independent components higher than the number of tablet pure compounds, and then higher than the physico-chemical rank of the hyperspectral dataset. This latter criterion can be defined as the number of variability sources in hyperspectral dataset, including both chemical and physical variations.

As far as a low dose compound is concerned, it can be assumed that its spectral and spatial contributions in the mixture dataset are low. The scarcity of the low dose compound can be associated with a low spectral variability in the hyperspectral dataset. If the spatial resolution is lower than the particle size of the low dose compound, its spectral information is mixed with the spectral contribution from the other actives and excipients. Moreover, because of its low spatial distribution, information of the low dose compound is supposed to be identified only in few

pixels. This spatial and spectral scarcity highlighted the challenge of extracting the associated pure signal by using a blind source separation method.

In this work, an over-fitted model using a number of components higher than the ICA_by_blocks method selection was used (15 components instead of 9). By using this high number of components, higher than the physico-chemical rank of the matrix, the spectral information from the low dose compound was extracted from the noise part of the matrix. The JADE algorithm was used to perform ICA decomposition. This algorithm starts by applying a singular value decomposition on the centered data with the objective of whitening and reducing the number of rows in the matrix (i.e. by calculating a scaled loadings matrix and a scaled scores matrix using singular value decomposition). If a model with *n* independent components is applied, *n* principal components are calculated in the whitening step (based on singular value decomposition). By reducing the dimensions of the initial matrix, some information can be lost, especially for a low dose compound. Therefore, this specific case requires a model with a sufficient number of components, which can be higher than the theoretical number of compounds in the spectral matrix or higher than the physico-chemical rank of hyperspectral dataset.

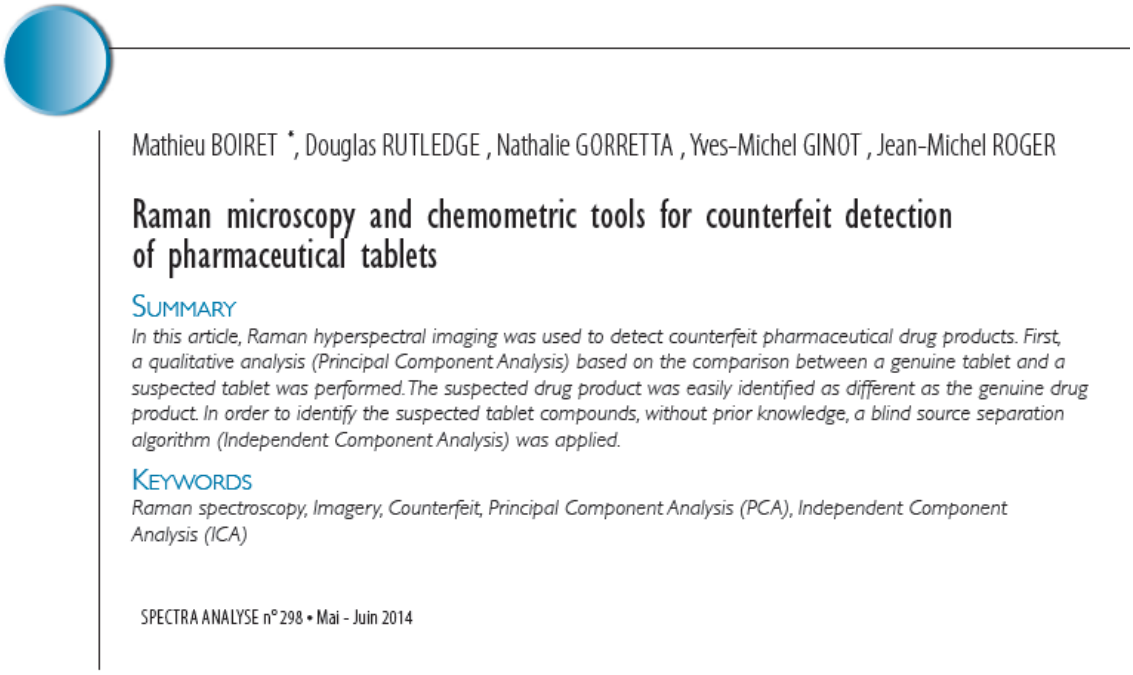
By projecting the pre-processed matrix on the calculated signal, distribution maps were easily displayed. Over-decomposed ICA model appeared as the only way to extract and detect the low dose compound, assuming that the compound pure spectrum is known. Without prior knowledge on the formulation, it would have been difficult, if not impossible, to identify a signal correlated to the magnesium stearate pure spectrum.

To conclude this first part, ICA was an interesting tool to extract pure spectra of actives and excipients in the studied formulation, without prior knowledge. The number of components, i.e. the number of signals, can be estimated by using ICA_by_blocks method but, due to spectral variability, it is often higher than the real number of compounds. In the case of a low dose compound, ICA_by_blocks method is not suitable and an over-fitted ICA model, with a number of components higher than the real number of products or higher than the physico-chemical rank of the matrix, must be used. One of the drawbacks of the ICA method was identified on the distribution maps assessment. Indeed, once the signals are calculated, distribution maps can be evaluated by projecting the initial spectral matrix on the signals. However, depending on the product contributions in the hyperspectral dataset, the distribution maps can be in accordance with the expected results or can be totally different, due to spectral correlations between drug products. In order to improve the distribution maps quality, and to enhance some spectral

variations, several pre-processing tools, such as derivatives, can be applied on the data before ICA decompositions.

In a future work, the previous results should be confirmed on different tablets, including low dose compounds which provide various Raman spectra, with different spectral responses or with different interactions between actives and excipients. Moreover, the JADE algorithm is one of the algorithms available for ICA decomposition. Among the different algorithms available, we have chosen to work with this algorithm because, in the chemometric community, it is the easiest method to understand and implement. It optimises the second and fourth order cumulants from the data and, although it is known to be slow for large data sets, it does not require any gradient searches and consequently avoids the convergence problems that sometimes occur with other algorithms [126]. JADE algorithm requires to reduce the matrix dimension and thus can lose the interesting information part linked to the low dose compound. In order to confirm the results of this chapter, it would be interesting to test other ICA algorithms, based on other decompositions, to evaluate the differences between these tools and to study advantages and drawbacks of each approach.

In addition and to expand these results, ICA was tested on a real case example for counterfeit sample analysis in **art. V** (Figure III-12). In this work, Raman hyperspectral imaging and PCA were firstly used to identify a counterfeit pharmaceutical drug product. In a second phase, hyperspectral dataset of the counterfeit sample was analysed with ICA, without prior knowledge on the formulation. The two main products in the tablet were easily identified: metformin (a well-known active different that the one used in the genuine formulation) and microcrystalline cellulose. This algorithm appeared as particularly powerful for this specific application where the studied formulation is never known.



Mathieu BOIRET *, Douglas RUTLEDGE , Nathalie GORRETTA , Yves-Michel GINOT , Jean-Michel ROGER

Raman microscopy and chemometric tools for counterfeit detection of pharmaceutical tablets

SUMMARY

In this article, Raman hyperspectral imaging was used to detect counterfeit pharmaceutical drug products. First, a qualitative analysis (Principal Component Analysis) based on the comparison between a genuine tablet and a suspected tablet was performed. The suspected drug product was easily identified as different as the genuine drug product. In order to identify the suspected tablet compounds, without prior knowledge, a blind source separation algorithm (Independent Component Analysis) was applied.

KEYWORDS

Raman spectroscopy, Imagery, Counterfeit, Principal Component Analysis (PCA), Independent Component Analysis (ICA)

SPECTRA ANALYSE n° 298 • Mai - Juin 2014

Figure III-12 Application of independent component analysis on counterfeit samples

Chapter IV: Use of multivariate curve resolution for identification of a low dose compound

1. Introduction 51

2. Materials and Methods 53

2.1. Samples 53

2.2. Raman imaging system 53

2.3. Pre-processing 54

2.4. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) 54

3. Results and discussion 56

3.1. Exploratory analysis 56

3.2. MCR-ALS 59

3.2.1. Non-negativity and local rank constraints 59

3.2.2. Effect of PCA filtering on MCR-ALS results 62

3.2.3. Pure spectrum augmented matrix 66

4. Conclusions 68

Preamble

As previously mentioned, the main objective of this thesis is the identification of low dose compound information (i.e. signal and distribution map) by using chemometric tools on hyperspectral dataset. In this work, a Raman hyperspectral image of a tablet including a low content product was used as experimental data. In the previous chapter, ICA was used without prior knowledge to extract pure signals and distribution maps. Satisfying results were obtained for the main excipients and actives. However, some limitations were highlighted for identification and distribution map assessment in the case of a low dose compound.

In this chapter, another well-known chemometric technique is tested and challenged: multivariate curve resolution-alternating least squares (MCR-ALS). Previously, this method has been successfully applied on hyperspectral dataset but to our knowledge, the detection of a low dose compound was not studied, especially from Raman hyperspectral dataset.

In this work, MCR-ALS is applied on the studied Raman data to identify the low dose compound and to provide its associated distribution in a tablet. Due to the low spectral variability of the compound (comparing with the other products of the formulation) and because MCR-ALS algorithm is based on variance decomposition, this objective appeared as a real challenge. Different approaches are proposed and compared (Figure IV-1), using initially filtered or non-filtered data, or using a column-wise augmented dataset before starting the MCR-ALS optimisation procedure including appended information on the low dose compound.

Note that this work has been performed with a Raman microscope and samples similar to those described in the chapter III. As this chapter is the reproduction of **Art. II** published in the Journal of Pharmaceutical and Biomedical Analysis in 2015, the readers will find some redundancies between chapter III and chapter IV in the materials and methods section.

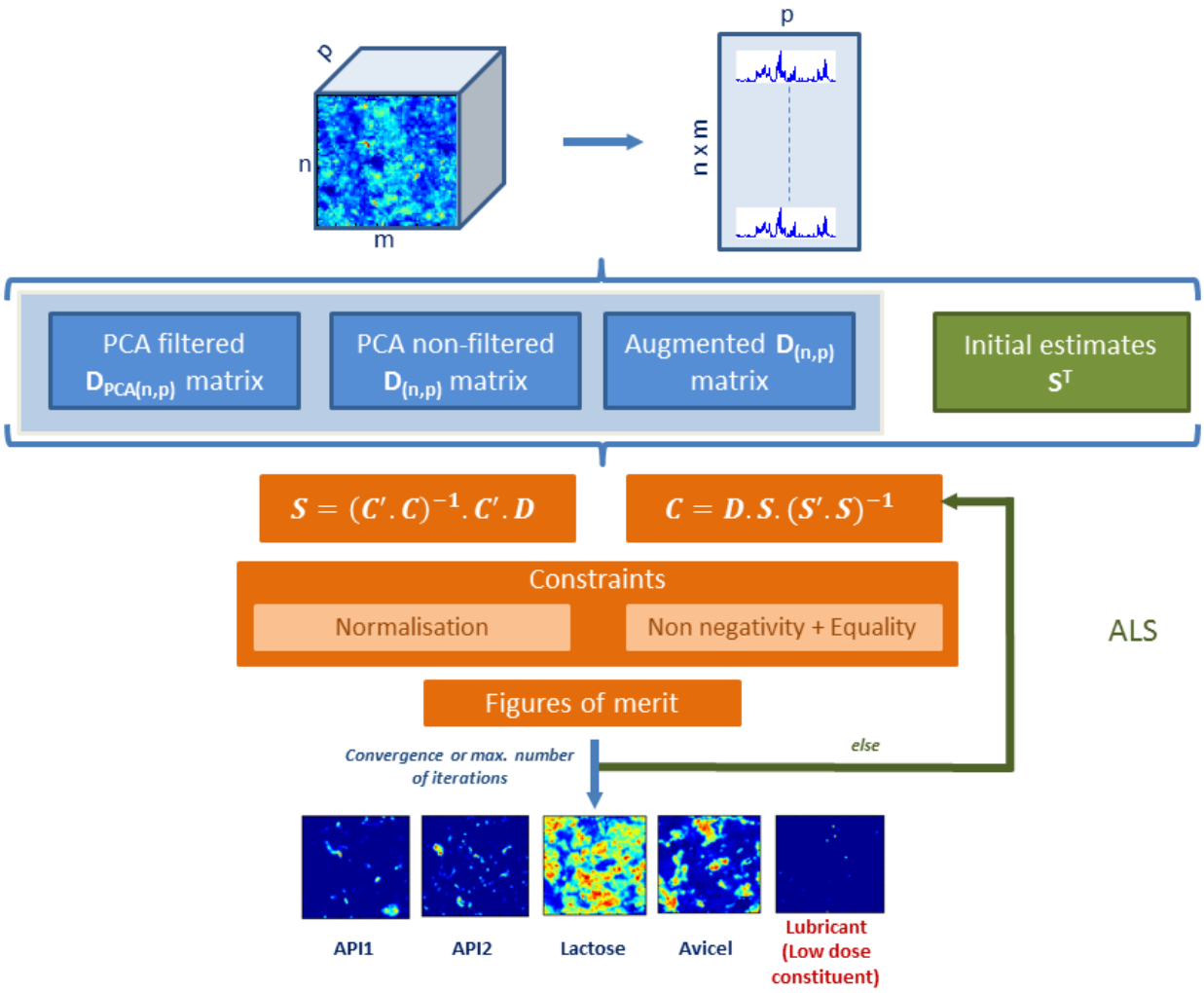


Figure IV-1 General scheme of the tested approaches in Chapter IV

DISTRIBUTION OF A LOW DOSE COMPOUND WITHIN PHARMACEUTICAL TABLET BY USING MULTIVARIATE CURVE RESOLUTION ON RAMAN HYPERSPECTRAL IMAGES[2]

1. Introduction

In the last decade, the use of imaging coupled with vibrational spectroscopies (near infrared, mid infrared, fluorescence and Raman) has grown quickly in research and development environments. The spatial and spectral information contained in hyperspectral images can be associated with the distribution of the different constituents within the sample. Different areas such as polymer research [127], biomedical analysis [128], environment field [108] and pharmaceutical development [129] are using these new analytical tools based on vibrational hyperspectral imaging. During the analytical lifecycle of a pharmaceutical drug product, hyperspectral imaging became a very powerful technique to explore the compound distributions on the tablet surface or within a powder mixture [130]. This technology appeared as innovative and promising to ensure the final quality of the drug product [131] from the development to the production.

Because of the huge amount of data contained in hyperspectral images, a direct interpretation of the acquired images is often not possible. Therefore, several chemometric tools have previously been applied [15; 132]. Qualitative analyses such as Principal Component Analysis (PCA) have already been used with near infrared [133] and Raman [134] chemical imaging in order to study the compound distribution in a sample. Since PCA is mainly linked to the dataset variability and as calculated loadings do not have chemical meaning, this approach is used as a descriptive method. To extract quantitative information at a global and pixel level, principal component regression (PCR) and partial least squares regression (PLS-R) have already demonstrated through several studies that they were powerful chemometric techniques [135; 136]. However, these methods can be time consuming and difficult to implement as they usually require a calibration step to develop predictive models. To overcome this problem, resolution methods seem to be a good alternative.

[2] Mathieu Boiret, Anna de Juan, Nathalie Gorretta, Yves-Michel Ginot, Jean-Michel Roger. **Distribution of low dose compound within pharmaceutical tablet by using multivariate curve resolution on Raman hyperspectral images**. *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 103 (2015) 35-43.

The aim of resolution methods is to provide the distribution maps and pure spectra related to the image constituents of a sample from the information contained in the raw image [137]. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is one of the most famous tools applied on hyperspectral images [83; 138]. MCR-ALS decomposes the initial data in a bilinear model, assuming that the observed spectra (i.e. each pixel of the image) are a linear combination of the spectra of the pure components in the system. In order to ensure an accurate resolution, constraints have to be used during the optimization process. Indeed, due to rotational or intensity ambiguities, resolution of a multicomponent hyperspectral image might not be unique [139]. Different constraints were established and tested [81; 140]. In image resolution, non-negativity, spectral normalization and local rank analysis are generally the most successful tools. Local rank analysis describes the spatial complexity of an image by identifying the rank of a pixel neighbourhood area. Combined with reference spectra of the image constituents, the absence of one or more specific constituent in a pixel can be highlighted. Some constraints used for the resolution of a chemical process, such as unimodality, closure or hard-modelling should not be used to analyse hyperspectral images because concentration profiles in the pixels of an image do not present the global continuous evolution that process profiles have [141].

Raman chemical imaging, because of its advantages such as negligible sample preparation, high chemical specificity and high spatial resolution, emerges as a new analytical tool in the quality control process of a solid drug product [142]. Final drug products are usually manufactured by using at least one active pharmaceutical ingredient (API) and several excipients. To improve powder flowability, most of the pharmaceutical manufacturing process includes a lubricant in the final drug formulation [143]. This compound is commonly present in a very low concentration in the powder blend and a spectroscopic bulk analysis will not be able to extract its contribution. Indeed, the corresponding variance of this constituent is very weak comparing with the other compounds of the sample. PCA, which aims at describing the directions of maximum global variance in the data, may have difficulties in retrieving information linked to a low dose constituent when the variance allocated to this component is similar in level to noise, which is often large in hyperspectral images. By offering the possibility to acquire images with a high spatial resolution, Raman chemical imaging coupled with appropriate chemometric methods appears as a promising technique to detect a low dose compound within a solid drug formulation.

In this work, MCR-ALS was applied on Raman chemical imaging data in order to provide the distribution of actives and excipients in a commercialised tablet. MCR-ALS was challenged by

trying to identify the low dose lubricant in the hyperspectral image. The effect of using algorithms driven by finding directions of maximum variance explained is studied. In this sense, the effect linked to the first step of noise-filtering based on PCA, which is often used in MCR-ALS to remove noise and non-useful spectral information, is studied. By applying MCR-ALS on a noise-filtered PCA matrix, it is shown that the information of the low dose constituent may be lost during data reduction. The comparison between the MCR-ALS decomposition on a filtered and a non-filtered PCA matrix is presented. Moreover, to keep the low dose constituent information during the PCA reduction, calculations are performed on an augmented matrix including the low dose constituent spectrum. The necessity of using appropriate pre-processing methods and constraints to find out the correct information linked to these low dose constituents is emphasized. This article shows the strategies to be followed in MCR-ALS analysis to retrieve correct information for low dose image constituents, from pre-processing, conditions to drive the iterative optimization to proper inclusion of constraints.

2. Materials and Methods

2.1. Samples

A commercial coated tablet of Bipreterax®, prescribed for arterial hypertension treatment and commercialised by "Les Laboratoires Servier", was used for the study. It is also known as Perindopril/Indapamide association. Final drug product contains respectively 4 mg of Perindopril (API1) and 1.25 mg of Indapamide (API2). Actives are known to have several solid state forms, but only one of them is present in this formulation. Major core excipients are lactose monohydrate, microcrystalline cellulose (Avicel). Magnesium stearate (MgSt), which is used as a lubricant, was added to the blend before compression with a theoretical mass concentration corresponding to 0.5% w/w. In order to analyse the tablet core, the coating was removed by eroding the sample with a Leica EM Rapid system (Leica, Wetzlar, Germany). A visual examination of the tablet did not provide any information concerning the distribution of the different compounds within the tablet.

2.2. Raman imaging system

The image was collected using a RM300 PerkinElmer system (PerkinElmer, Waltham, MA) and the Spectrum Image version 6.1 software. The microscope was coupled to the spectrometer and spectra were acquired through it with a spatial resolution of 10 μm in a Raman diffuse reflection mode. Wavenumber range was 3200–100 cm^{-1} with a resolution of 2 cm^{-1}. Spectra were

acquired at a single point on the sample, then the sample was moved and another spectrum was taken. This process was repeated until spectra of points covering the region of interest were obtained. A 785 nm laser with a power of 400 mW was used. Two scans of 2 s were accumulated for each spectrum. An image of 70 pixels per 70 pixels corresponding to 4900 spectra was acquired for a surface of 700 µm by 700 µm.

2.3. Pre-processing

Data were preprocessed in order to remove non-chemical biases from the spectra (scattering effect due to non-homogeneity of the surface, interference from external light source, spikes due to cosmic rays, random noise). First of all, data were spike-corrected in order to reduce the effect of cosmic rays [61]. The spectral range was reduced in order to focus only on the region of interest, corresponding to a Raman shift from 1800 cm^{-1} to 200 cm^{-1}. Reduced spectra were preprocessed by asymmetric least squares (AsLS) to correct baseline variations due to fluorescence contributions [68]. Finally, to enhance slight spectral variations, a Savitzky-Golay first derivative with a 2^{nd} order polynomial smoothing on a 9 points window [59] was applied.

2.4. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)

A brief description of the MCR-ALS algorithm is given here. The algorithm was previously described in detail in Refs. [81; 140]. As any resolution methods, the main goal of MCR-ALS is decomposing the original matrix $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$ (n samples or rows and p variables or columns) of a multi-component system into the underlying bilinear model which assumes that the observed spectra are a linear combination of the spectra of the pure components in the system:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad \textbf{(IV-1)}$$

where **C** is the matrix of concentration profiles, $\mathbf{S}^{\mathbf{T}}$ the matrix of pure responses (i.e. spectra) and **E** contains the experimental error. In resolution of spectroscopic images, $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$ is the matrix of the unfolded image, **C** contains the concentration profiles that, conveniently refolded, show the distribution maps of each image constituent and $\mathbf{S}^{\mathbf{T}}$ contains the associated pure spectra [144].

In order to provide chemically meaningful profiles (i.e., pure spectra and distribution maps) and to reduce intensity and rotational ambiguities in the MCR solutions, constraints must be properly chosen during the iterative MCR-ALS process. Since concentrations of the constituents

should not be negative, a non-negativity constraint was applied. Moreover, the calculated spectral profiles in matrix $\mathbf{S^T}$ were normalized at each iteration. To identify where the constituents of the drug product are present or absent in the image, the Fixe Size Moving Window Evolving Factor Analysis (FSMW-EFA) method was applied to the data [145]. This method provides the local complexity of a sample by performing singular value decomposition by moving a window of pixels across the full image. A window contains a specified number of spectra (at least 4, corresponding to a specific pixel and its neighbours). By calculating singular value maps of the sample, the presence of overlapped compounds in a pixel area can be displayed. By selecting a specific threshold, a corresponding local rank map can be provided by plotting the number of significant singular values above the threshold. This approach, due to its local character, is particularly well adapted to identify a compound with a low signal or with a low concentration within the sample because small local areas are analyzed one at a time. By comparing the local rank information with reference spectral information, missing constituents on particular pixels can be known [141].

Figures of merit of the optimization procedure are the lack of fit (lof) and the explained variance (R^2). The lack of fit is used to check if the experimental data were well fitted by the MCR-ALS procedure. These two criteria are calculated as follow:

$$lof(\%) = 100\sqrt{\frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} D_{i,j}^2}} \qquad \textbf{(IV-2)}$$

$$R^2 = \frac{\sum_{i,j} D_{i,j}^2 - \sum_{i,j} e_{i,j}^2}{\sum_{i,j} D_{i,j}^2} \qquad \textbf{(IV-3)}$$

where $\mathbf{D_{i,j}}$ is the input element of the original matrix $\mathbf{D_{(n,p)}}$ and $\mathbf{e_{i,j}}$ the related residual element after using the MCR-ALS model (see equation IV-1). Input element can be the original element from $\mathbf{D_{(n,p)}}$ or the element of a noise filtered PCA matrix $\mathbf{D_{PCA_{(n,p)}}}$ using the same number of components as in the MCR-ALS. A noise filtered PCA matrix $\mathbf{D_{PCA_{(n,p)}}}$ can be obtained as follows:

$$\mathbf{D_{PCA_{(n,p)}}} = \mathbf{U_{(n,k)}S_{(k,k)}V^T_{(k,p)}} \qquad \textbf{(IV-4)}$$

where $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{V^T}$ are calculated by singular value decomposition of the original $\mathbf{D_{(n,p)}}$ matrix and k is the number of the known constituents in the drug product. The PCA reduced matrix

corresponds to a filtered matrix in a reduced space. This matrix should contain the major part of the spectral variance without noise.

MCR-ALS must be initialised by a first estimate of **C** or $\mathbf{S}^{\mathbf{T}}$ matrix. Initial estimates are generally obtained by purest variable selection methods, such as SIMPLISMA (Simple-to-use Interactive Self-Modelling Mixture Analysis) [87]. This method identifies the most dissimilar spectra (or sample) in the dataset. However, due to the homogeneity of a pharmaceutical sample, it could be difficult to identify a pure pixel corresponding to a single constituent. Most of the time, the theoretical formulation of the sample is known during the development process. So pure reference spectra acquired with the same spectrometer and the same acquisition parameters can be selected as initial estimates to start the optimisation process.

In this article, three approaches will be tested and discussed in order to display the distribution of actives and excipients, including the low dose constituent. The first approach starts with the noise filtered PCA matrix $\mathbf{D}_{\mathbf{PCA}_{(\mathbf{n},\mathbf{p})}}$ calculated from equation (IV-4) using a component number *k* equal to the theoretical number of constituents in the formulation. The second approach consists of increasing the number of components to generate the noise filtered PCA matrix, from *k* to the maximum number of variables, the latter meaning working with the raw non-filtered data set. The third approach consists of using an augmented matrix, where the information of the low dose constituent is added, ensuring the extraction of its contribution during the noise filtering step.

3. Results and discussion

3.1. Exploratory analysis

Because of the spectral variability, applying multivariate data analysis on raw data would not lead to accurate results. Spectra were preprocessed in order to remove baseline variations and cosmic rays. A spike correction algorithm and asymmetric least squares were applied. In order to enhance low variations, a Savitzky-Golay first derivative with a window size of 9 points and a 2nd polynomial order was calculated (Figure IV-2).

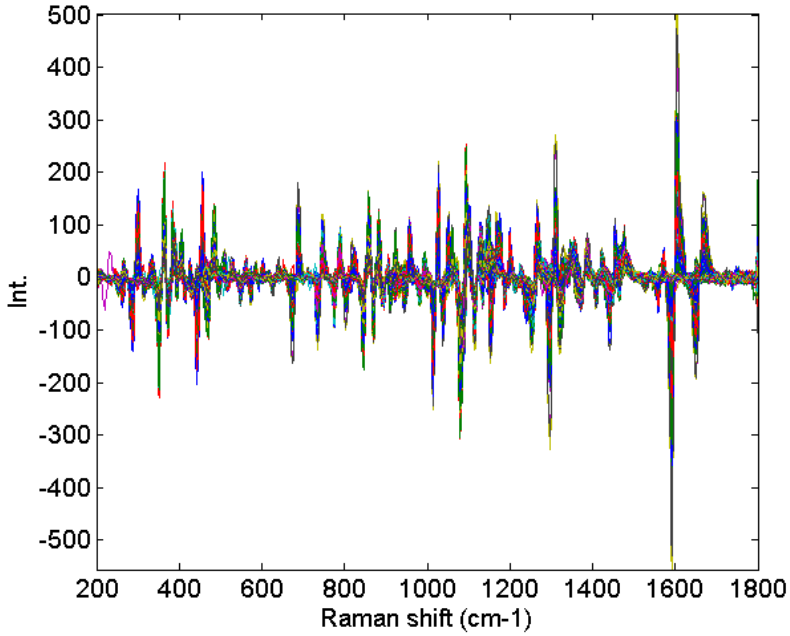


Figure IV-2 Preprocessed Raman spectra (AsLS and first derivative)

By observing the mean intensity plot of the image (mean intensity in each pixel), no useful information about compound distributions was extracted (results not shown). Therefore, chemometric tools have to be used in order to extract meaningful distributions of the different compounds. As a descriptive method, PCA was applied on the preprocessed data. By calculating appropriate principal components, that describe the maximum variance of the data set and are orthogonal to each other, PCA decomposes the preprocessed matrix in scores (related to distribution maps) and loadings (related to spectra) matrices [49; 146]. Figure IV-3 shows the image scores results of the five first principal components. Different distributions and agglomerates were highlighted in the images. In this particular example, by knowing the studied formulation and by observing the calculated loading vectors, the distribution maps of PC1 and PC5 were linked to the lactose variability, while distribution maps of PC2, PC3 and PC4 were respectively linked to the distributions of API1, avicel and API2.

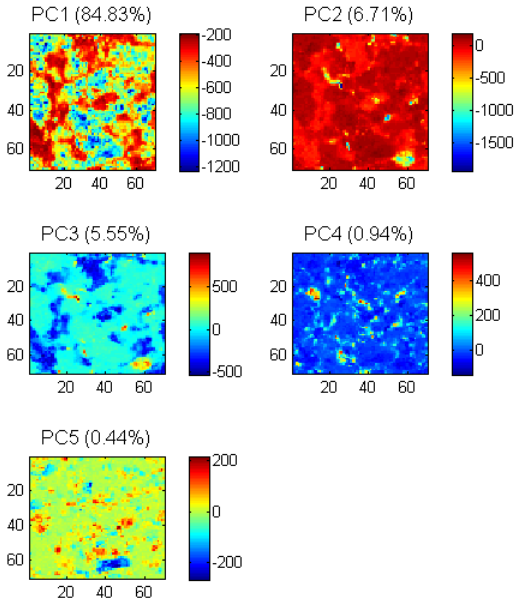


Figure IV-3 PCA scores: five first components associated with their explained variances. Different distributions and agglomerates were highlighted. PC1 and PC5 were linked to the lactose variability, while PC2, PC3 and PC4 were respectively linked to the distributions of API1, avicel and API2.

Even if PCA analysis provides a first approximation of the component distribution within the sample, the contribution of magnesium stearate was not extracted with this approach. By observing the cumulative variance explained by the PCA model, it was shown that 98.5% of the variance was captured with 5 components, which means that 1.5% of the spectral variability was not explained by the model. From PC6, the variance contained in the principal components was lower than 0.2% of the total variance and reached a plateau of 0.02% of variance explained per component, which could be associated with a non-structured noise contained in the spectral matrix.

Theoretical spectral variance Var_i of the magnesium stearate was estimated to 0.5% of the total variance and was calculated by using the following equation:

$$Var_i = 100 \times \frac{\sum_{i,j}\left(c_i s_i^T\right)^2}{\sum_{i,j}\left(CS^T\right)^2} \qquad \textbf{(IV-5)}$$

where **C** and $\mathbf{S}^T$ are respectively the theoretical concentrations and the pure reference spectra of each constituent i. Due to the low concentration of magnesium stearate within the drug product, and because of the homogeneity aspect of the powder mixture before compression, the spectral variance of the lubricant might be lower or higher than 0.5%, depending on the studied area of the tablet.

Several hypotheses could explain the non-identification of magnesium stearate within the spectral matrix. Due to its low concentration, the lubricant could either be present on a limited number of pixels or could either be missing in the studied area. The associated spectral information could have led to overlapped features with other components or could have been spread into noise contributions.

PCA is mainly linked to the variability contained within the hyperspectral dataset, expressed as a combination of orthogonal components. Even if it provides a first approximation of the four major constituent distributions, the low spectral variability linked to the lubricant was not displayed on the five first components. Moreover, due to their unclear chemical meaning, loadings are difficult to interpret. To overcome this issue, MCR-ALS algorithm and appropriate constraints were used to enhance the chemical information of the decomposition.

3.2. MCR-ALS

3.2.1. Non-negativity and local rank constraints

MCR-ALS was initialized by using reference spectra of the five different constituents. Spectra were acquired with the same system and with the same parameters as the image. Image pre-processing tools were applied on the reference spectra (see section 2.3). To reduce rotational and intensity ambiguities, non-negativity and equality constraints were applied on the calculated concentrations. Lof and R^2 values were calculated according to equations (IV-2) and (IV-3).

By analysing the image locally, FSMW-EFA provides an estimation of the local complexity of the image [145]. Local rank map was obtained by calculating singular value decomposition on a 4 pixel window moving across the whole data. In general, the number of pixels has to be equal or higher than the total number of the image constituents but in this case, due to the high spatial resolution, the hypothesis was advanced that the five compounds could not be present in the same pixel. Four eigenvalues were calculated for each pixel group. Each component singular

values were sorted in increasing order (Figure IV-4). By choosing an appropriate threshold which separates significant singular values from noise, the local rank map was displayed (Figure IV-5). (Note that the threshold is selected visually, based on the fact that singular values associated with noise are very small and similar among them and lay at the bottom of plot in Figure IV-4). The number of missing components for a specific pixel was calculated by removing the local rank value of the pixel to the total rank of the matrix (chosen as the number of theoretical constituents). By calculating correlation coefficients between the raw pixel spectrum and each of the reference spectra, the constituent with the lowest correlation was identified as absent. The absence of a particular component in a pixel was not confirmed unless the correlation coefficient between the pixel spectrum and the reference spectrum of that component is equal or smaller than the largest element in the correlation matrix for that particular component. Results were afterwards encoded in an absence matrix $\mathbf{C}_{sel}$ (Figure IV-6) containing null values in the concentration elements of the missing components and "not-a-number" (NaN) values in other pixels (unconstrained pixels) [141].

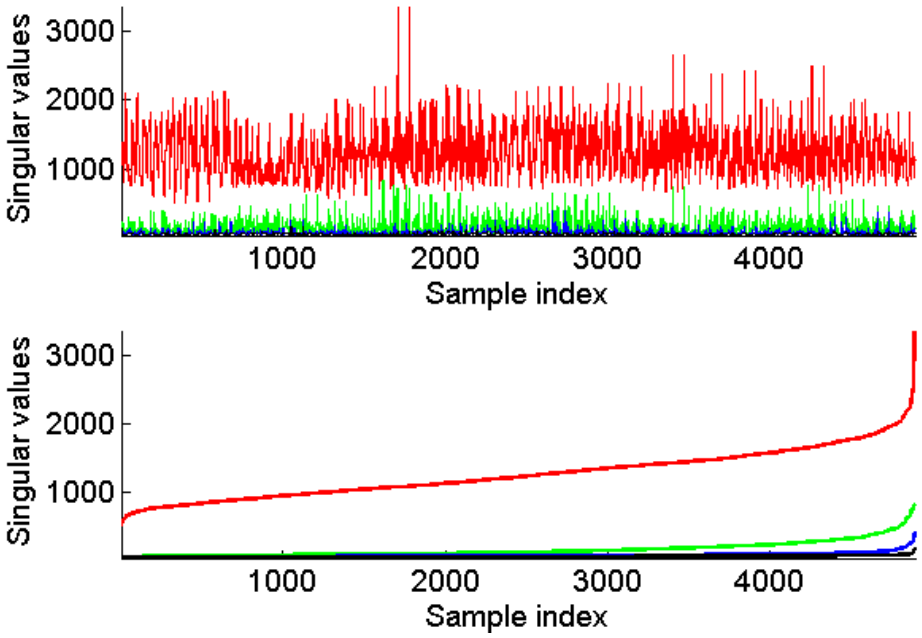


Figure IV-4 Singular values plot (top: non-sorted singular values, bottom: sorted singular values)

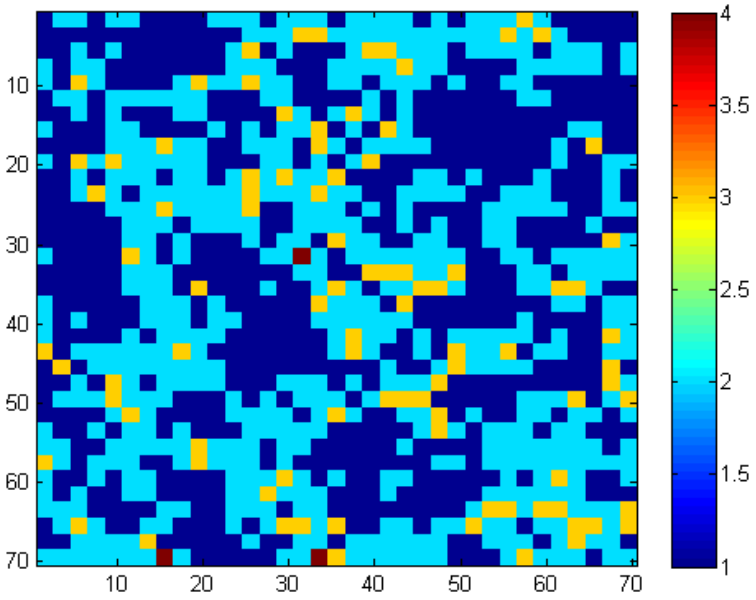


Figure IV-5 Local rank map obtained by choosing an appropriate threshold which separates significant singular values from noise.

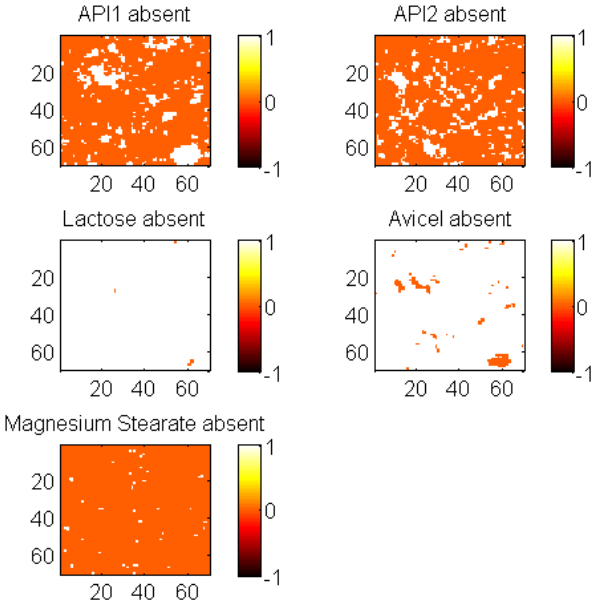


Figure IV-6 C_{sel} matrix (Orange: absence of the constituent, White: presence of the constituent)

3.2.2. Effect of PCA filtering on MCR-ALS results

In all cases, MCR-ALS was applied on the preprocessed data by using the constraints previously described (see section 3.2.1). The initial preprocessed matrix was reduced (noise-filtering) by using the five first vectors of the PCA decomposition of $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$. MCR-ALS on the filtered PCA matrix provides an optimum value after 9 iterations. 97.9% of the variance was explained with a lack of fit calculated on the initial $\mathbf{D}_{(\mathbf{n},\mathbf{p})}$ and the reduced $\mathbf{D}_{\mathbf{PCA}(\mathbf{n},\mathbf{p})}$ matrices respectively equal to 14.7 and 7.9. Correlation coefficients between calculated spectra and reference spectra were displayed in Table IV-1. The four first calculated spectra were highly correlated to the two actives and the two major excipients whereas the fifth component was not correlated to the magnesium stearate or to other constituents.

	API1	**API2**	**Lactose**	**Cellulose**	**MgSt**
$\mathbf{S_{opt}.1}$	**0,98**	0,16	0,02	-0,02	-0,05
$\mathbf{S_{opt}.2}$	0,15	**0,97**	0,04	0,05	0,07
$\mathbf{S_{opt}.3}$	0,10	0,01	**0,99**	0,14	0,03
$\mathbf{S_{opt}.4}$	0,00	0,05	0,15	**0,95**	-0,03
$\mathbf{S_{opt}.5}$	0,02	0,21	-0,09	0,07	0,08

Table IV-1 Correlations between MCR-ALS calculated S_{opt} and the reference spectra (PCA filtered dataset)

By starting MCR-ALS after a PCA reduction of the data, the magnesium stearate contribution was associated with the non-explained variance. In our example, the theoretical number of components in the drug product is equal to 5. The matrix $\mathbf{D}_{\mathbf{PCA}(\mathbf{n},\mathbf{p})}$ calculated by equation (IV-4), is then calculated by using the five first components of the PCA decomposition. With 5 components, 98.5% of the total variance was explained, which means that 1.5% of the variance was not included in the iterative MCR-ALS process. This part of the non-explained variance contains essentially noise but, due to the low concentration of magnesium stearate, could also contain the spectral contribution of this constituent.

In order to improve the MCR-ALS results and to extract magnesium stearate contribution, MCR-ALS analysis on a PCA-filtered matrix including progressively a larger number of principal components was tested. MCR-ALS decomposition was performed by using a PCA-filtered $\mathbf{D}_{\mathbf{PCA}(\mathbf{n},\mathbf{p})}$ matrix using an increasing number of components, from 5 to the total number of variables. For the first iteration, the $\mathbf{D}_{\mathbf{PCA}(\mathbf{n},\mathbf{p})}$ matrix was built by using the five first vector of the

PCA reduction. The following MCR-ALS calculation was performed by adding an additional principal component to calculate the $\mathbf{D_{PCA_{(n,p)}}}$ matrix. This process was repeated until the number of principal components was equal to the number of variables, corresponding to the use of the preprocessed non-filtered initial $\mathbf{D_{(n,p)}}$ matrix. For each MCR-ALS decomposition from 5 to 100 components, the highest correlation coefficient between the resolved spectra and the pure reference spectrum of magnesium stearate is displayed (Figure IV-7).

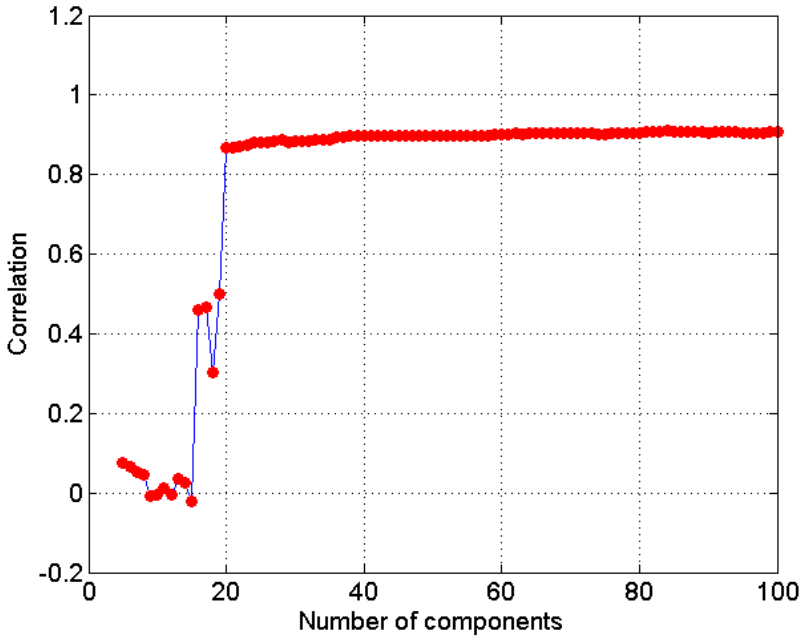


Figure IV-7 Highest correlation between the calculated spectra (S_{opt}) and the reference spectrum of magnesium stearate (for each iteration of a PCA filtered matrix built from 5 to 100 components)

By using less than 20 principal components to reproduce the $\mathbf{D_{PCA_{(n,p)}}}$ matrix, the contribution of magnesium stearate was not extracted. Using 20, the correlation between the calculated spectrum and the reference magnesium stearate spectrum was equal to 0.87 and reached 0.90 after a using 50 principal components to reproduce the matrix. As it is shown in Table IV-2, where MCR-ALS was applied on a $\mathbf{D_{PCA_{(n,p)}}}$ built with k = 5, 10, 15, 20, 50, the results of the two active principal ingredients and the two major excipients were not modified. In this case, using less than 20 components to build the matrix $\mathbf{D_{PCA_{(n,p)}}}$ lose the magnesium stearate contribution.

Number of principal components used to reproduce the $D_{PCA_{(n,p)}}$ matrix	**5**	**10**	**15**	**20**	**50**
Iterations	9	5	5	3	3
R^2	99.4	98.8	98.7	98.6	98.4
Lof (%)	7.9	11.12	11.6	11.9	12.8
Cor. $S_{opt}1$/API1	0.98	0.98	0.98	0.98	0.98
Cor. $S_{opt}2$/API2	0.97	0.97	0.97	0.97	0.97
Cor. $S_{opt}3$/lactose	0.99	0.99	0.99	0.99	0.99
Cor. $S_{opt}4$/cellulose	0.95	0.95	0.95	0.95	0.95
Cor. $S_{opt}5$/MgSt	0.08	-0.01	-0.02	**0.87**	**0.90**

Table IV-2 MCR-ALS results according to the number of components used to build the PCA reduced $D_{PCA_{(n,p)}}$ matrix

In order to keep the maximum information, the initial preprocessed $\mathbf{D_{(n,p)}}$ matrix (i.e. the PCA non-filtered dataset) was used to start the iterative MCR-ALS process. Non-negativity and local rank constraints on concentrations were applied. The optimum was reached after 3 iterations, with a lack of fit equal to 14.7 and a percentage of variance explained equal to 97.8.

Correlations between calculated spectra and API1, AP12, lactose and avicel were respectively equal to 0.98, 0.97, 0.99 and 0.95. Distributions and contributions of the different constituents were then displayed in Figure IV-8. Major excipients (lactose and cellulose) are identified across the whole image in distribution maps 3 and 4. Agglomerates of API 1 and API 2 were highlighted in the top left and right distribution maps. The correlation between the calculated spectrum and the magnesium stearate reference was equal to 0.90 (Figure IV-9). By using a non-filtered PCA matrix with appropriate constraints, the information linked to the low dose constituent was extracted. The non-filtering option can be the choice when there are no references that can indicate in an objective manner the number of PCs necessary to include a minor constituent. As shown in Figure IV-8, only few pixels of the image contained the lubricant ($\mathbf{C_{opt.5}}$), which could be explained by its low concentration within the drug product.

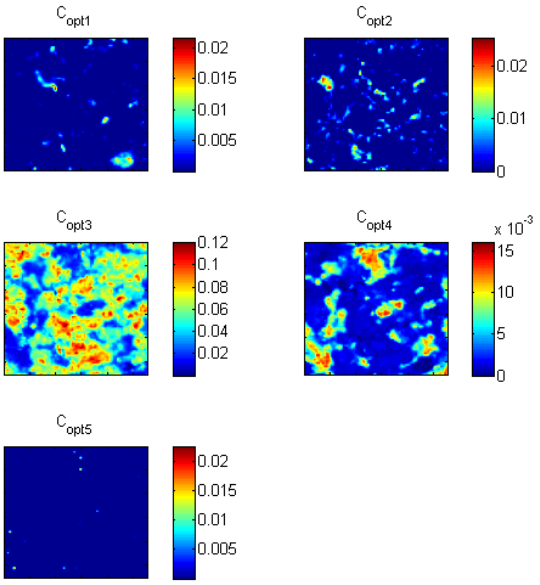


Figure IV-8 Distribution maps of drug substance constituents (PCA non-filtered dataset)

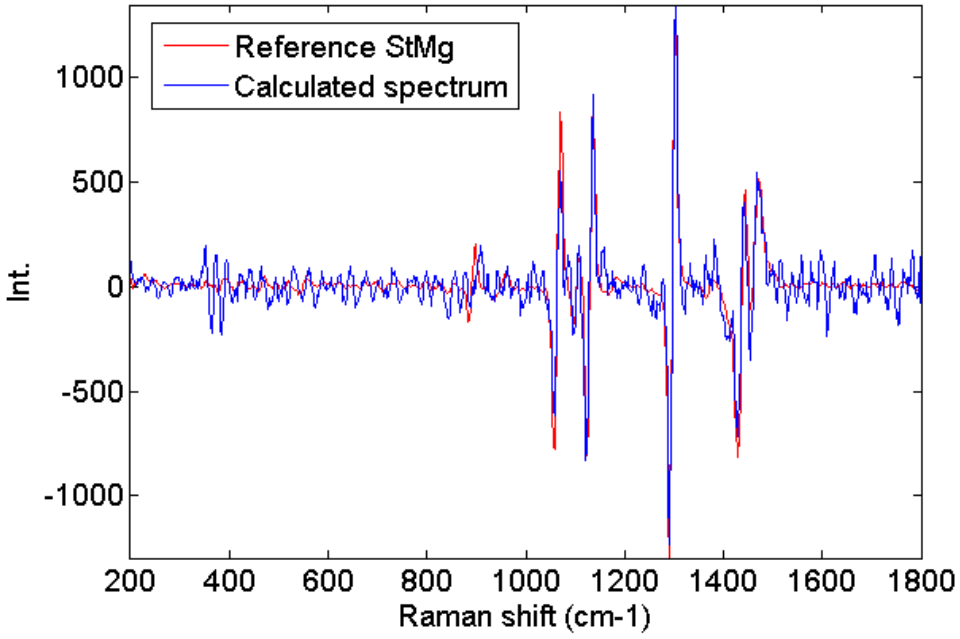


Figure IV-9 S_{opt} versus reference spectrum of magnesium stearate

3.2.3. Pure spectrum augmented matrix

The preprocessed data matrix was column-wise augmented to form a multiset structure including the magnesium stearate preprocessed pure spectrum [147]. For this type of matrix augmentation, the bilinear model can be written as:

$$\begin{pmatrix}\mathbf{D_1}\\ \mathbf{D_2}\end{pmatrix} = \begin{pmatrix}\mathbf{C_1}\\ \mathbf{C_2}\end{pmatrix}.\mathbf{S^T} + \begin{pmatrix}\mathbf{E_1}\\ \mathbf{E_2}\end{pmatrix} = \mathbf{C_{augm}}.\mathbf{S^T} + \mathbf{E_{augm}} \qquad \textbf{(IV-6)}$$

where $\mathbf{S^T}$ is the pure spectral matrix of the different compounds present in the considered preprocessed $\mathbf{D_1}$ data matrix and the augmented $\mathbf{D_2}$ pure spectrum matrix. In these two matrices, the chemical compounds have to be the same, but their concentration profiles can be different. Non negativity of concentration and local rank constraints were applied on the data as it was described in section 3.2.1. In multiset analysis, a new constraint based on correspondence among species can be used. This constraint fixes the presence or absence of components in concentration matrix, always taking into account the sequence of components in the initial estimates to encode the information on presence/absence correctly. This presence or absence information is coded in binary format and introduced into the MCR algorithm. For $\mathbf{D_1}$, the correspondence among species vector was fixed to [1, 1, 1, 1, 1] as each constituent was supposed to be in the drug product whereas, for $\mathbf{D_2}$, only one value corresponding to the lubricant was fixed to 1, corresponding to the vector [0, 0, 0, 0, 1] (Note that this code is valid as long as MgSt is the fifth profile in the spectral estimates used in the MCR analysis). When a particular component is not present in a concentration matrix, the elements in the related profile are set to zero. This type of constraint contributes significantly to the elimination of rotational ambiguities.

By adding information of the low dose constituent in the matrix, the PCA reduction of the multiset provides a different model, which ensures the extraction of the lubricant information. The MCR-ALS can then be performed as usual, by using a first step of PCA reduction with 5 components. The optimum was reached after 6 iterations, with a lack of fit equal to 8.3 (with respect to $\mathbf{D_{PCA_{(n,p)}}}$) and 16.1 (with respect to $\mathbf{D_{(n,p)}}$) and a percentage of variance explained equal to 97.4%.

Correlations between calculated MCR-ALS $\mathbf{S_{opt}}$ spectra and the five reference spectra were respectively equal to 0.98, 0.96, 0.99, 0.95 and 0.99 (Table IV-3) which ensure an appropriate resolution of the studied system. In Figure IV-10, distributions of API1, API2 and the two main

excipients were in accordance with the previous results obtained from MCR-ALS on a PCA filtered or non-filtered dataset. However, because of the high correlation between the calculated $\mathbf{S_{opt.5}}$ spectrum and the magnesium stearate reference spectrum, the distribution of the lubricant can be easily observed in the $\mathbf{C_{opt.5}}$ distribution map. As for the PCA non-filtered approach, only few pixels were highlighted with the lubricant contribution, which could be explain by its low concentration within the drug product.

	API1	API2	Lactose	Cellulose	MgSt
$\mathbf{S_{opt}.1}$	**0,98**	0,16	0,02	-0,02	0.03
$\mathbf{S_{opt}.2}$	0,16	**0,96**	0,04	0,06	0,19
$\mathbf{S_{opt}.3}$	0,08	0,04	**0,99**	0,14	-0,09
$\mathbf{S_{opt}.4}$	-0,10	0,07	0,16	**0,95**	-0,20
$\mathbf{S_{opt}.5}$	0,02	0,20	-0,09	0,07	**0.99**

Table IV-3 Correlations between MCR-ALS S_{opt} and the reference spectra (column-wise augmented dataset)

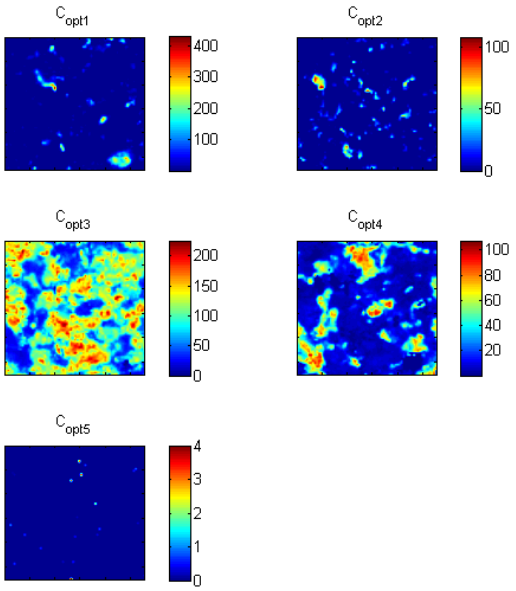


Figure IV-10 Distribution maps of drug substance constituents (augmented matrix approach)

4. Conclusions

MCR-ALS was applied on Raman Chemical images in order to study the distribution of actives and excipients within a pharmaceutical drug product. This article was focused on the identification of a low dose constituent within a formulation. Three different approaches were tested. First, MCR-ALS was performed on a PCA reduced dataset built by using a number of components equal to the number of constituents within the formulation. Due to the low spectral variability of the lubricant, the PCA reduction did not extract the corresponding information and the MCR-ALS process was not able to find out this product. However, distribution of actives and major excipients were in accordance with the known formulation. In order to ensure the conservation of the low dose constituent contribution within the dataset, a sequential PCA reduction process was tested. For each iteration, a new PCA reduced dataset was generated (from 5 to 100 components) and used for MCR-ALS calculations. It was shown that the lubricant information was not present in the iterative MCR-ALS process unless 20 components were used. From a PCA non-filtered dataset, the magnesium stearate distribution was detected by using appropriate non-negativity and local rank constraint. Results showed the distribution of the five constituents with high correlations between the calculated signals and the pure reference spectra. Finally, the initial preprocessed dataset was column-wise augmented with magnesium stearate preprocessed pure spectrum. By using a correspondence among species constraint properly defined, the PCA reduction of the matrix kept the lubricant information and then, the decomposition of the Raman chemical image provided high correlated calculated spectra with reference and well-defined actives and excipients distribution map.

This study demonstrates the ability of MCR-ALS to extract the contribution of a low constituent of a solid drug product from Raman hyperspectral images. The choice of appropriate pre-processing methods, constraints, data structures used and modus operandi was important to reach the objective. Raman Chemical images, known as a useful tool to study the distribution of compounds in a solid drug product, might be used to study the distribution of low dose constituents as a lubricant, an impurity or a crystalline form transformation.

Contributions of chapter IV

In this chapter, we focused on the strategies to be followed in MCR-ALS analysis to retrieve correct information for low dose image constituents, from pre-processing, conditions to drive the iterative optimization to proper inclusion of constraints. MCR-ALS is a well-known chemometric technique for data analysis on vibrational spectroscopy dataset. It was previously applied on hyperspectral imaging in order to study the distribution of actives and excipients. In our work, we tested the MCR-ALS ability to extract information from a low dose compound in a mixture dataset. As MCR-ALS usually starts with a noise filtering step by applying PCA on the studied dataset, we made the hypothesis that the information from a low dose compound could be lost before starting the iterative process.

First, pre-processed data were observed by applying a non-centered PCA. It is generally associated with a singular value decomposition on the variance-covariance matrix (the center of the matrix is the mean value of the dataset), but, in the case of a non-centered matrix, the variance-covariance matrix should be viewed as a scatter matrix (the center of the matrix is equal to zero). This decomposition provides the direction of the maximum variability across the data. By applying a non-centered PCA on the studied matrix, and by knowing the studied formulation, the distribution maps of the main compounds were highlighted. However, the contribution of magnesium stearate, the low dose lubricant, was not extracted with this methodology. PCA pointed out some difficulties in retrieving information linked to this compound because the variance allocated to this constituent can be considered as similar in level to noise. Indeed, with a theoretical concentration of 0.5% (w/w) in the drug product, the spectral variance contained in the data is weak. By using a principal PCA-filtered dataset as a first step of the MCR-ALS approach, the lubricant information is scattered in the non-explained variance (in the non-structured noise part) and the associated distribution in the tablet cannot be highlighted.

Two hypotheses could explain the non-identification of magnesium stearate within the spectral matrix. First, due to its low concentration, the low dose compound could be missing in the studied area. However, according to equation II-11, a number of spectra equal to 4900 (70 pixels x 70 pixels) should be sufficient to detect the lubricant (probability = 0.999). Second hypothesis could be explained with spectral information which could have led to overlapped features with other components or could have been spread into noise contributions.

Considering the MCR-ALS procedure to study the compound distributions, three different approaches were tested and challenged to extract the targeted low dose information. The first started the iterative process with the noise PCA-filtered matrix using a number of components equal to the theoretical number of compounds in the formulation. The second consisted of increasing the number of components to generate a noise PCA-filtered matrix, using a number of components from *k* to the maximum number of variables, the latter meaning working with the raw non-filtered dataset. The third approach consisted of using an augmented matrix, where the spectral information of the low compound was added before calculation, ensuring the extraction of its contribution during the noise filtering step.

It was shown that a sufficient number of components to generate the PCA-filtered matrix must be used in order to keep the lubricant variability within the dataset or, otherwise, work with the raw non-filtered data. Different models were built using an increasing number of components to perform the PCA reduction. It was shown that the magnesium stearate information can be extracted from a PCA model using a minimum of 20 components. In the last part, a column-wise augmented matrix, including a reference spectrum of the lubricant, was used before starting MCR-ALS process. PCA reduction was performed on the augmented matrix, to ensure that the magnesium stearate contribution was included within the MCR-ALS calculations. By using an appropriate PCA reduction, with a sufficient number of components, or by using an augmented dataset including appended information on the low dose component, the distribution of the two actives, the two main excipients and the low dose lubricant were correctly recovered.

Moreover, the effect of using appropriate pre-processing methods and constraints to find out the correct information linked to the low dose compound was assessed. Spike correction, baseline correction, and derivative were used for two reasons: i/ to avoid harmful signal contributions and ii/ to enhance slight spectral variations. Because initial estimates of concentrations or spectra are needed to start the MCR-ALS iterative process, pure spectra of each compound were acquired using the same experimental conditions.

In order to improve the resolution and to reduce rotational and intensity ambiguities, equality constraint was applied on the calculated concentrations. This constraint was calculated by estimating the local rank of each image pixel and by providing a map of absence/presence for each drug compound. This approach used a threshold on singular values in order to separate significant values from noise. But, this threshold was selected visually, based on the fact that singular values associated with noise are very small and similar among them. Without knowing the sample or the distribution of a low dose constituent, it can be difficult to select the threshold.

The studied dataset was previously used in **Art. I** and we already had a good idea concerning the distribution of magnesium stearate in the image, which helped us to choose the proper threshold. However, it is important to notice that a modification of the threshold value can have a huge impact on the resolution, and then can lead to incorrect results.

This latter observation led us to the following chapter where we will focus on the optimization of the equality constraint by proposing a new method of mapping based on orthogonal projections. Because the methodology based on the calculation of singular value on moving windows appeared as uncertain and risky in the case of a low dose product, we will propose an alternative method which could provide suitable absence/presence maps.

Chapter V: An alternative method for presence/absence maps determination by orthogonal projections

1. Introduction 75

2. Theory 77

2.1. Notations 77

2.2. Pretreatment using orthogonal projections 77

2.3. Multivariate curve resolution-alternating least squares (MCR-ALS) 78

2.4. Proposed approach to determine presence/absence maps of compounds to set local rank constraints 79

3. Materials and methods 82

3.1. Raman microscopy 82

3.2. Samples 82

3.2.1. Simulated data 82

3.2.2. Real dataset 85

4. Results and discussion 85

4.1. Principal component analysis (PCA) on pure images 85

4.2. Proposed approach on simulated data 86

4.3. Proposed approach on real dataset 90

5. Conclusions 93

Preamble

In the previous chapter, MCR-ALS was used to provide the distribution of a low dose compound in a pharmaceutical drug product. Different approaches were tested and proposed to keep the low dose information before starting the iterative process. Avoid the PCA-filtering step appeared as essential to obtain a good resolution. Moreover, constraints on spectra and concentrations must be applied between iterations in order to reduce ambiguities. At the end of chapter IV, some limitations about the absence/presence maps (used as equality constraint) calculation were highlighted.

Equality constraint is known to be a powerful tool which allows the identification of presence or absence of a specific compound. However, in the case where the low dose compound has low spectral contributions, two issues may arise by using a method based on singular value decomposition. First, singular value decomposition might encounter some difficulties to extract the variability linked to the low dose compound if the spectral response is low, because the associated variance is weak. Second, comparison between image spectra and reference pure spectra might be difficult to perform.

In this work, we propose an alternative procedure to set the presence/absence maps of compounds for the determination of MCR-ALS equality constraint. In order to focus on the useful information from a compound, the proposed approach is based on orthogonal projection to a space containing the contributions to be removed, i.e. the interference subspace which contains environmental, acquisition or physical variations from the other compounds. By working within the signal space, describing the P-dimensional space in which the observations can be represented as vectors, it ensures the detection of a compound without requiring important variations between samples.

The proposed approach will be firstly tested on a simulated dataset. By knowing pure spectra and distribution of actives and excipients, the method capability to provide maps of absence/presence will be assessed. In a second part, the proposed approach will be tested on a real tablet image. Absence/presence maps and MCR-ALS results will be presented and discussed.

Note that this work has been performed with a Raman microscope and formulation similar to those described in the chapter III and chapter IV. As this chapter is the reproduction of **Art. III** published in Analytica Chimica Acta in 2015, the readers will find some redundancies between chapters III/IV and chapter V in the materials and methods section.

SETTING LOCAL RANK CONSTRAINTS BY ORTHOGONAL PROJECTIONS FOR IMAGE RESOLUTION ANALYSIS: APPLICATION TO THE DETERMINATION OF A LOW DOSE PHARMACEUTICAL COMPOUND[3]

1. Introduction

The use of imaging coupled with vibrational spectroscopies has shown a huge interest in research and development environments [15], especially to control the drug product quality during development and beyond post-marketing authorisation [49]. It provides spatial and spectral information associated with the distribution of the different compounds within the sample. Direct interpretation of the acquired images is often not possible and several chemometric tools have previously been published to aid in this task [132]. Qualitative analyses such as principal component analysis (PCA) [134] or independent component analysis (ICA) [89] have already been used as a descriptive method to study compound distributions in a sample by Raman chemical imaging. To extract quantitative information at a global and local pixel level, principal component regression (PCR) and partial least squares regression (PLS-R) have been shown to be powerful chemometric techniques [135]. However, these methods can be time consuming and difficult to implement since they require a calibration step to develop predictive models.

By avoiding the calibration step, resolution methods were identified as a good alternative to study the compound distribution within a pharmaceutical drug product. They provide the distribution maps and pure spectra related to the image compounds of a sample from the information contained in the raw image [148]. Multivariate curve resolution-alternating least squares (MCR-ALS) has been used on Raman hyperspectral images to study the distribution of actives and excipients [83; 138]. In order to ensure an accurate resolution, constraints have to be used during the optimization process. In image resolution, non-negativity, spectral normalization and local rank analysis are generally the most successful constraints [84]. Local rank analysis describes the spatial complexity of an image by identifying the rank of a pixel neighbourhood area. Combined with reference spectra of the image compounds, the absence of one or more specific compound in a pixel can be highlighted. To identify where the compounds of the drug product are present or absent in the image, the fixed size image window evolving factor analysis (FSIW-EFA) method can be applied to the data [145]. This method provides the

[3] Mathieu Boiret, Anna de Juan, Nathalie Gorretta, Yves-Michel Ginot, Jean-Michel Roger. **Setting local rank constraints by orthogonal projections for image resolution analysis: Application to the determination of a low dose compound**. *Analytica Chimica Acta* (2015), Vol. 892 (2015) 49-58

local complexity of a sample by performing singular value decomposition by moving a window of neighbouring pixels across the full image. By comparing the local rank information with reference spectral information, missing compounds on particular pixels can be known. The local complexity and the correct definition of the presence/absence maps are relevant steps of the MCR-ALS algorithm. Indeed, the quality in the resolution of the system depends on the adequacy and correct setting of constraints. If pure compound pixels or pixels with absent compounds are present in the images, both singular value decomposition and identification of missing compounds should led to the identification of the presence or absence of the studied compound and hence should help to provide better MCR-ALS results, less affected by ambiguity.

In the case of a low dose product, it can be assumed that the compound is not homogenously distributed (i.e. it is present in a few pixels at low concentrations). Spatial and spectral information is scarce because only few pixels of the image contain the product of interest and the associated variances are mixed with the other compounds of the formulation. In order to keep the maximum of information during the iterative process, MCR-ALS has to be performed without PCA-based filtering matrix [149] with the appropriate constraints. In cases where the low dose compound has additionally a low spectral response, two problems may arise to obtain proper local rank maps and related maps of presence/absence for this kind of compounds. Firstly, singular value decomposition applied on moving window might encounter some difficulties to extract the variability linked to the low dose compound if the spectral response is low, since the associated variance is weak. Second, since the correlation between pure spectra of the formulation is not null (i.e. spectra are not orthogonal), construction of presence/absence maps, based on the comparison between image spectra and reference pure spectra might be difficult to set up especially if the contribution of the signal of the low dose compound to the pixel spectrum measured is low.

Previous works have been published on the detection of a low dose compound by vibrational spectroscopy within a pharmaceutical drug product [96; 97; 100] and some of them focused on the detection limit of the analytical method [101]. The net analyte signal (NAS) concept was used in the pharmaceutical environment to improve the spectral interpretability [150] of model results. It was defined as the part of the signal that is orthogonal to the spectra of the other components [151]. For one component of interest, two definitions were proposed [102; 103]. NAS was first defined as “the part of the spectrum of the component of interest that is orthogonal to the spectra of the other components”. Afterwards, the definition evolved into “the part of the raw signal that is useful for prediction of the component of interest”. NAS has a conceptual meaning and is very difficult to measure. It can be viewed as a particular case of

preprocessing methods based on orthogonal signal projection approaches [152-154]. The use of NAS pretreatment appeared as an interesting tool to accurately resolve the analyte signal of a low dose compound and allow the construction of a quantitative model [104]. Several adaptations of these approaches can be considered, depending on the spectral basis (i.e. space containing the contributions to be removed) used for projecting the original dataset.

In this article, we propose an alternative procedure to set the presence/absence maps of compounds for later use as local rank constraints. The proposed approach is based on orthogonal projection to a space containing the contributions to be removed (i.e. the interference subspace). Each compound has its proper subspace containing spectral variability due to the environment, acquisition or physical variations. This variability can be viewed as a basis of vectors with an appropriate set of dimensions to build the interference subspace. By orthogonally projecting a spectrum to this basis, interferences are removed and only information of the compound of interest is kept. Since the method is not based on variance decomposition, it should be well adapted for a drug product which contains a low dose compound located in few pixels and with a low spectral response. Spectral comparison between the projected spectrum and a pure projected spectrum of the compound of interest can lead to the presence/absence maps used as local rank constraints in the MCR-ALS iterative process.

2. Theory

2.1. Notations

Vectors are noted in bold lowercase, matrices in bold uppercase, and scalars in italic lowercase characters. Vectors are arranged in lines and one line represents one spectrum. The transposed forms of a vector $\mathbf{x}$ and a matrix $\mathbf{X}$ are noted $\mathbf{x}^{\mathbf{T}}$ and $\mathbf{X}^{\mathbf{T}}$, respectively. $\mathbf{I}$ is the identity matrix of dimensions *p x p,* where *p* is the number of variables in a spectrum. $\mathbf{x}$ and $\mathbf{X}$ orthogonally projected to a detrimental basis $\mathbf{K}$ are noted $\mathbf{x}_{\perp}$ and $\mathbf{X}_{\perp}$. $\mathbf{\Sigma}$ is the Euclidian orthogonal projector to $\mathbf{K}$.

2.2. Pretreatment using orthogonal projections

Orthogonal projections can be applied as a preprocessing method by orthogonally projecting spectra to a basis of detrimental information or interferences. This idea was previously illustrated by the concept of NAS which was defined in the literature as the part of the sample spectrum that is related to the analyte and orthogonal to the interferences [151]. The performance of the pretreatment is directly explained by its ability to obtain a good

approximation of a basis **K** including the detrimental information. The basis **K** can be set up by using pure spectra or information extracted from experimental design or from models or by using calibration datasets.

Let **K** be a basis of detrimental information of dimensions *l x p* (*l* spectra/signals, *p* variables), including all the chemical information of the formulation, except the one of the compound of interest. The Euclidian orthogonal projector to **K** can be calculated by applying:

$$\mathbf{\Sigma} = \mathbf{I} - \mathbf{K^T}\left(\mathbf{K}\,\mathbf{K^T}\right)^{-1}\mathbf{K} \qquad \textbf{(V-1)}$$

Assuming that $\mathbf{x_n}$ is a spectrum of dimension *1 x p* from the matrix **X** of dimension *n x p*, a spectrum $\mathbf{x_{n\perp}}$ is obtained after a projection of $\mathbf{x_n}$ orthogonally to **K** through the Euclidian orthogonal projector **Σ**:

$$\mathbf{x_{n\perp}} = \mathbf{x_n}\,\mathbf{\Sigma} \qquad \textbf{(V-2)}$$

For all spectra, the projected matrix $\mathbf{X_\perp}$ orthogonal to **K** can be obtained by applying:

$$\mathbf{X_\perp} = \mathbf{X}\left(\mathbf{I} - \mathbf{K^T}\left(\mathbf{K}\,\mathbf{K^T}\right)^{-1}\mathbf{K}\right) \qquad \textbf{(V-3)}$$

2.3. Multivariate curve resolution-alternating least squares (MCR-ALS)

The algorithm was previously described in detail in Refs. [81; 82]. The main goal of MCR-ALS is decomposing the original matrix $\mathbf{X_{(n,p)}}$ (*n* samples or rows and *p* variables or columns) of a multicomponent system into the underlying bilinear model which assumes that the observed spectra are a linear combination of the spectra of the pure components in the system:

$$\mathbf{X} = \mathbf{C}\,\mathbf{S^T} + \mathbf{E} \qquad \textbf{(V-4)}$$

where **C** is the matrix of concentration profiles, $\mathbf{S^T}$ the matrix of pure responses and **E** contains the experimental error. In resolution of spectroscopic images, $\mathbf{X_{(n,p)}}$ is the matrix of the unfolded image, **C** contains the concentration profiles that, conveniently refolded, show the distribution maps of each image constituent and $\mathbf{S^T}$ contains the associated pure spectra [144]. To keep the maximum of information during the iterative process, and especially when a low dose compound is studied, MCR-ALS must be performed on the $\mathbf{X_{(n,p)}}$ matrix without PCA-based filtering [149]. In order to provide chemically meaningful profiles (i.e. pure spectra and

distribution maps) and to reduce intensity and rotational ambiguities in the MCR solutions, constraints must be properly chosen during the iterative MCR-ALS process. Since concentrations of the compound and Raman intensities should not be negative, a non-negativity constraint (n-n constraint) is often applied.

To exploit the presence or absence of a compound in image pixels, local rank constraints can be used. A previous approach, based on FSIW-EFA, was performed by applying local singular value decompositions [27]. This method was identified as a powerful tool by providing local information on pixels. Coupled with reference spectral information, it can provide a map of absence for each constituent. This method can be applied in all instances and it is the only option when there is no prior information on the composition of the samples and, hence, interference spaces as proposed in section 2.2 can not be designed. However, in the case of a low dose compound with a low spectral signal, the identification of the presence and absence of this compound might be difficult. In this work, an alternative method to the FSIW-EFA method is proposed. This approach, based on orthogonal projections, is particularly adapted to the identification of a low dose compound with low intensity spectral signal within a pharmaceutical drug product and will be presented in the next section. Lack of fit and explained variance, two common figures of merit [82] of the MCR-ALS optimization procedure, were used to assess the model efficiency.

2.4. Proposed approach to determine presence/absence maps of compounds to set local rank constraints

In this work, the orthogonal projection pretreatment was used to set up presence/absence maps used as local rank constraints during MCR-ALS iterations. The basis **K** with information of the interference spaces was estimated by applying singular value decomposition on the suitable pure compound Raman images. Each pure Raman image is formed by several spectra, including all environmental, acquisition, physical and chemical variability for a compound. In order to include all the spectral variability in the interference space, an appropriate number of dimensions was chosen for each subspace **k** of the basis **K**. The final **K** basis can, therefore, have a rank higher than the number of compounds of the image to include all spectral variability that can be found associated with a single compound.

For each compound of interest *c*, let $\mathbf{K_{-c}}$ be the interference matrix, including all the variability of the drug compounds, except the information from *c*. The orthogonal projector to $\mathbf{K_{-c}}$ can be calculated as follow:

$$\boldsymbol{\Sigma_{-c}} = \mathbf{I} - \mathbf{K}_{-\mathbf{c}}^{\mathbf{T}}(\mathbf{K}_{-\mathbf{c}}\mathbf{K}_{-\mathbf{c}}^{\mathbf{T}})^{-1}\mathbf{K}_{-\mathbf{c}} \qquad \textbf{(V-7)}$$

By orthogonally projecting the spectra $\mathbf{x}_n$ of an image of the pharmaceutical formulation and a pure spectrum $\mathbf{s}$ of the compound of interest to the basis $\boldsymbol{\Sigma_{-c}}$, projected image spectra $\mathbf{x}_{n\perp}$ and a pure projected spectrum $\mathbf{s}_\perp$ can be calculated by:

$$\mathbf{x}_{n\perp} = \mathbf{x}_n\boldsymbol{\Sigma_{-c}} \qquad \textbf{(V-8)}$$

and

$$\mathbf{s}_\perp = \mathbf{s}\boldsymbol{\Sigma_{-c}} \qquad \textbf{(V-9)}$$

In this work, the mean spectrum of a pure compound image was used as the pure spectrum $\mathbf{s}$. In Figure V-1, inspired by [154], the basis $\mathbf{K_{-c}}$ is constituted of three subspaces corresponding to a formulation which contains four compounds. Each subspace is a set of loading vectors calculated by applying a non-centered PCA based on singular value decomposition on pure compound images (respectively 4, 3 and 2 loading vectors for interference spaces of compounds 1, 2 and 3). $\mathbf{x}_{n\perp}$ and $\mathbf{s}_\perp$ are calculated by orthogonally projecting an image spectrum and the pure spectrum of the product of interest to the detrimental subspace, i.e. the basis $\mathbf{K_{-c}}$. It is crucial to project the pure spectrum of each compound also onto the related interference space to keep only as a reference the orthogonal part of the pure spectra for later correlation studies. Performing the correlation between the orthogonal projection of any pixel spectrum and the raw pure spectrum of the pure compounds would lead to misleading conclusions because any pure compound spectrum is partially correlated to the rest of compounds in the system.

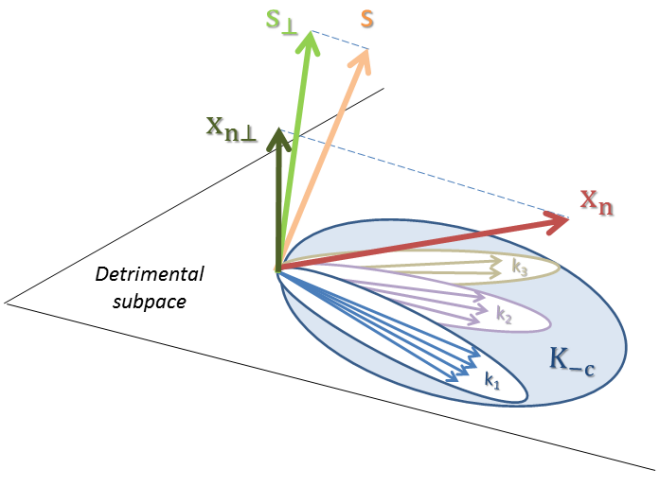


Figure V-1 Graphical representation of the proposed approach

For a compound, the presence/absence map is calculated by estimating the Pearson's correlation between the orthogonal projections of the pixel spectra $\mathbf{x}_{n\perp}$ and the pure spectrum $\mathbf{s}_{\perp}$:

$$\mathbf{r}_{(\mathbf{x}_{n\perp},\mathbf{s}_{\perp})_i} = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(\mathbf{x}_{n\perp i}-\overline{\mathbf{x}_{n\perp}})(\mathbf{s}_{\perp}-\overline{\mathbf{s}_{\perp}})}{\mathbf{S}_{\mathbf{x}_{n\perp i}}\mathbf{S}_{\mathbf{s}_{\perp}}} \qquad \textbf{(V-10)}$$

A high $\mathbf{r}_{(\mathbf{x}_{n\perp},\mathbf{s}_{\perp})_i}$ value corresponds to a spectrum strongly correlated to the pure projected spectrum $\mathbf{s}_{\perp}$. If the value is higher than a specified threshold, then the compound is considered as present. On the other hand, a low value will be associated with the compound absence. Selection of the threshold value can be considered as a critical parameter in the proposed approach. An appropriate value, ensuring the proper presence or absence of a compound, has to be selected by observing the correlations $\mathbf{r}_{(\mathbf{x}_{n\perp},\mathbf{s}_{\perp})_i}$ and the associated projected image spectra and pure projected spectrum. Absences of compounds in different pixels will be encoded and used as local rank constraints during the alternating least squares iterative process.

3. Materials and methods

3.1. Raman microscopy

Images were collected using a RM300 PerkinElmer system (Perkin Elmer, Waltham, MA) and the Spectrum Image version 6.1 software. The microscope was coupled to the spectrometer and spectra were acquired with a spatial resolution of 10µm in a Raman diffuse reflection mode. A 785nm laser with a power of 400mW was used. Wavenumber range was 3200–100cm^{-1} with a resolution of 2cm^{-1}. Spectra were acquired at a single point on the sample, then the sample was moved and another spectrum was taken. This process was repeated until spectra of points covering the region of interest were obtained.

3.2. Samples

3.2.1. Simulated data

A hyperspectral Raman image was synthesized by using pure compound images of lactose, avicel® (i.e. microcrystalline cellulose), indapamide (used as active pharmaceutical ingredient or API) and magnesium stearate. Pure tablets were prepared with a manual tablet press and Raman images of 40 pixels per 40 pixels corresponding to 1600 spectra were acquired for a surface of 400µm by 400µm. Two scans of two seconds were accumulated for each spectrum. In order to simulate a hyperspectral Raman dataset with various known concentrations of API and excipients, an image structure was manually designed by defining 8 different classes including circular and rectangular shapes (Figure V-2). Each class was constituted of different amount of the four pure products (Table V-1). In order to simulate concentration variability, concentrations were normally distributed around the target value (distribution was set up with a mean of zero and a standard deviation of one), leading to four various concentration maps (Figure V-3). The magnesium stearate, usually added in a pharmaceutical formulation as a lubricant, was used as the low dose compound and was only present in classes seven and eight with a target concentration of 5% w/w. Among the 1600 pixels, only six contained magnesium stearate. The spectral and spatial variability directly linked to the lubricant is weak and the distribution study of this compound appeared as a real challenge.

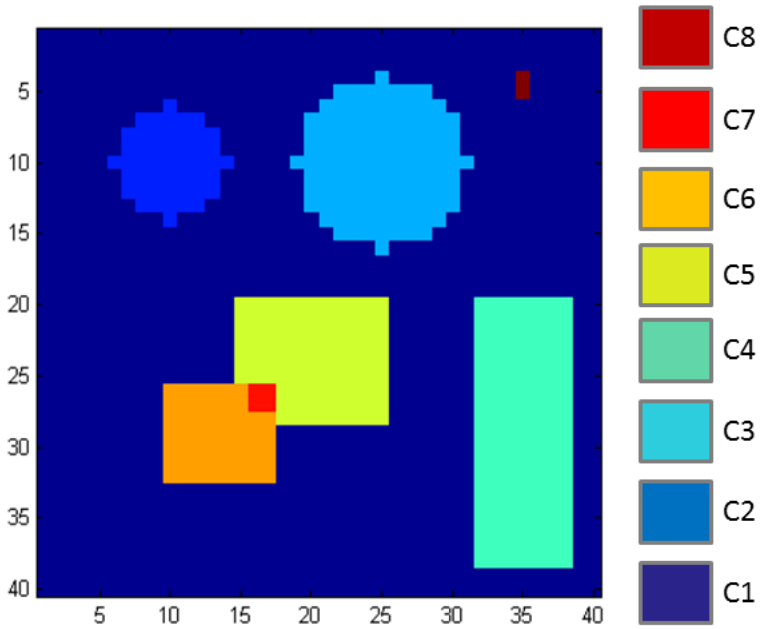


Figure V-2 Image distribution patterns used to build synthetic image (eight classes represented by eight different colours)

Classes	Lactose	Avicel®	MgSt	API
1	70	20	0	10
2	20	70	0	10
3	20	30	0	50
4	20	60	0	20
5	30	10	0	60
6	10	70	0	20
7	20	35	5	40
8	65	20	5	10

Table V-1 Target concentrations of the eight classes

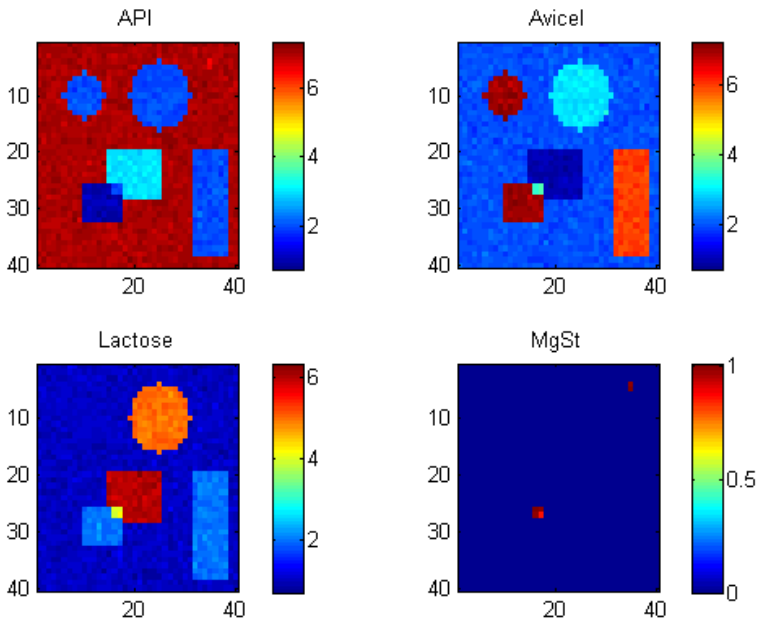


Figure V-3 Simulated distribution maps of lactose, avicel®, API and magnesium stearate (MgSt)

Each spectrum $\mathbf{[x_1, x_{2,\ldots}, x_i]}$ of the simulated image **X** of dimensions *n x m x p* where *n* means pixels in x- direction, *m* pixels in y- direction and *p* spectral channels was the result of the sum of the images of the pure compounds, conveniently modified to reflect the pixel concentration patterns assigned to the classes. For one spectrum *i*, a spectrum $\mathbf{x_i}$ was obtained as:

$$\mathbf{x_i} = \sum_{j=1}^{j} \mathbf{C_{i,j}S_{i,j}} \qquad \textbf{(V-11)}$$

where $\mathbf{C_{i,j}}$ is the concentration profile of the spectrum *i* that includes the concentration pattern of the compound *j* in the different classes and $\mathbf{S_{i,j}}$ the pure spectra (Figure V-4). In this way, the whole simulated image was obtained by using each of the 1600 spectra of the images of pure compounds. Using this simulation procedure, concentration variability was included within the final dataset. Moreover, due to the spectral variability included within the pure compound images, noise was also included in the hyperspectral data.

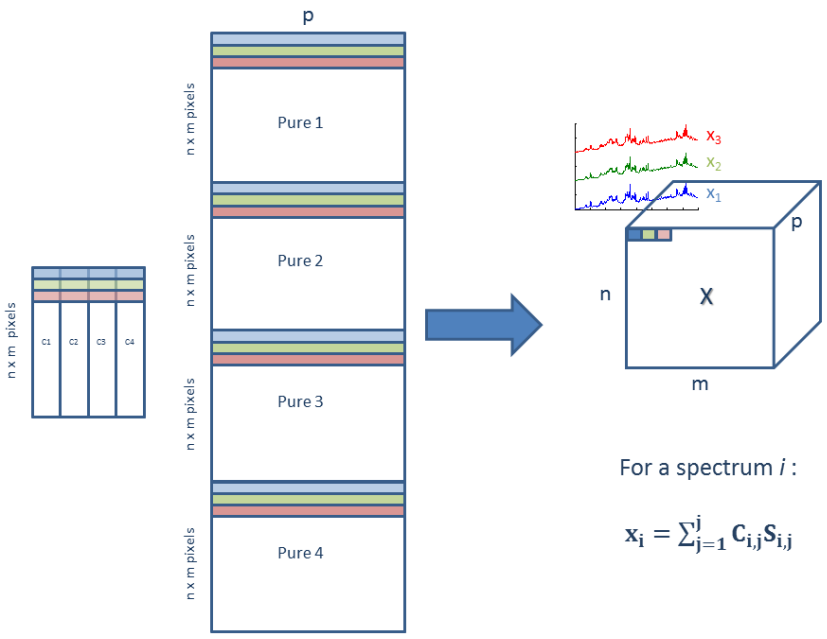


Figure V-4 Building of simulated data

3.2.2. Real dataset

A commercial coated tablet of Bipreterax® was used for the study. Bipreterax® is used for arterial hypertension treatment and is commercialised by "Les Laboratoires Servier". It is also known as Perindopril (active principal ingredient 2 or API 2) / Indapamide (active principal ingredient 1 or API 1) association and contains respectively 4mg of API 2 and 1.25mg of API 1 in the commercial drugs. Actives are known to have several solid state forms, but only one of them is present in this formulation. Major core excipients are lactose monohydrate and avicel® and minor excipient is magnesium stearate. In order to analyse the tablet core, the coating was removed by eroding the sample with a Leica EM Rapid system (Leica, Wetzlar, Germany). A visual examination of the tablet did not provide any information concerning the distribution of the different compounds within the tablet. Two scans of two seconds were accumulated for each spectrum. An image of 70 pixels per 70 pixels was acquired for a surface of 700µm by 700µm.

4. Results and discussion

4.1. Principal component analysis (PCA) on pure images

Due to physical and environmental variability such as particle size, polymorphism, water content and acquisition variation, spectral differences are often observed in a hyperspectral image of a pure compound. On Raman spectra, this variability can be illustrated by multiplicative

and additive effects due to scattering and fluorescence contributions. By applying non-centered PCA on pure Raman images, a well-known chemometric method based on singular value decomposition, the variability subspace of each compound can be identified. A reference image of Indapamide, of dimensions 40 pixels per 40 pixels was acquired. In Figure V-5, spectral variations can be observed between the 1600 spectra. Scores images and loadings, based on variance decomposition of the whole matrix, highlighted spectral variability included in the Raman image. The number of components of the model was chosen by observing the total variance explained.

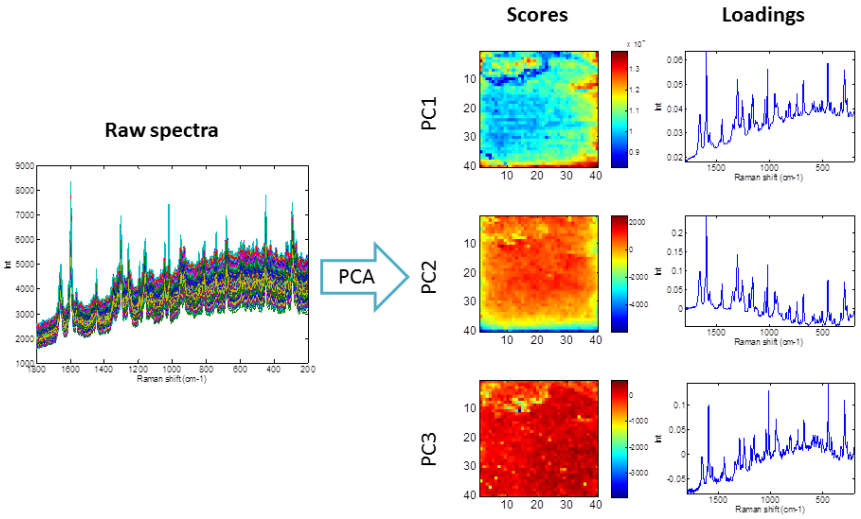


Figure V-5 Non-centered PCA on pure compound image of API. Raw spectra, scores maps and loadings

4.2. Proposed approach on simulated data

A non-centered PCA was performed on the four pure unfolded images in order to set up the interference basis. The number of components was chosen by explaining 99.9% of the total spectral variance. The interference basis $\mathbf{K}_{-\mathbf{c}}$ for lactose, avicel®, API and magnesium stearate were respectively built with 5, 5, 3 and 4 components (i.e. loading vectors).

Orthogonal projected spectra of the simulated image to the suitable interference space for each compound were displayed in Figure V-6. For lactose, avicel® and API, orthogonal projected signals were similar to the expected pure compound Raman spectra. For magnesium stearate,

Raman spectra orthogonally projected to the basis $\mathbf{K}_{-\mathbf{c}}$ including spectral variability and information of the three main formulation compounds (lactose, avicel® and API) provided noisy signal. Due to the weak spatial and spectral presence of magnesium stearate within the simulated formulation, direct interpretation of these projected spectra was not possible. Each projected spectral matrix was compared with the orthogonal projection of the related pure spectrum by correlation. Correlation maps displayed in Figure V-7 showed high correlations between the projected spectra of lactose, avicel® and API and the associated pure projected spectrum. The 6 pixels containing low concentrations of magnesium stearate provided correlations with the projected pure spectrum from 0.5 to 0.8. Other pixels of the image were not correlated to the projected pure spectrum of magnesium stearate.

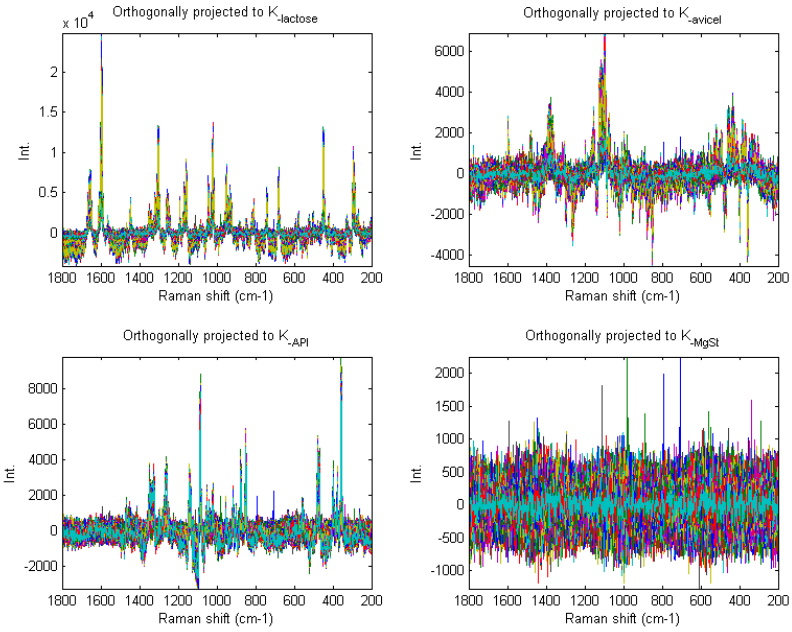


Figure V-6 Orthogonal projected spectra of the simulated data to the suitable interference space $\mathbf{K}_{-\mathbf{c}}$ for each compound

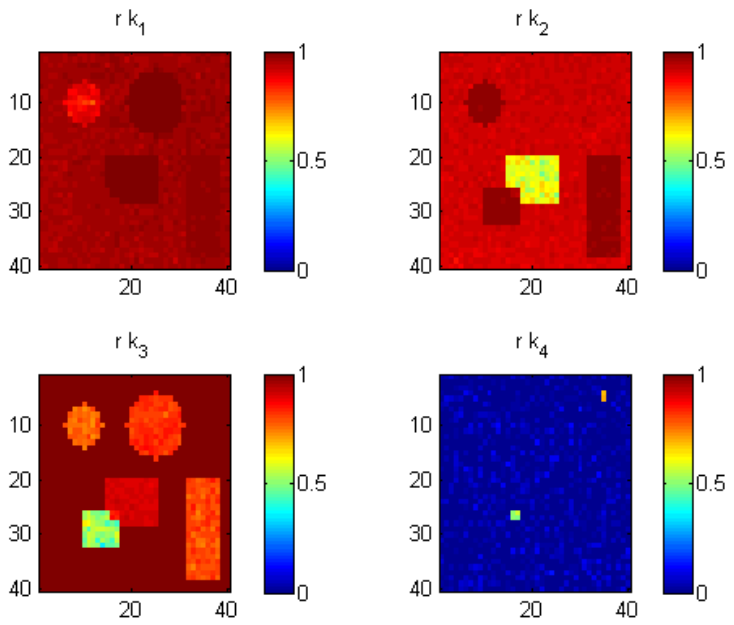


Figure V-7 Correlation maps $r_{(x_{n\perp},s_{\perp})_i}$ (k1 = lactose basis, k2 = avicel basis, k3 = API basis, k4 = MgSt basis)

By selecting a threshold which ensures the detection of a compound in a pixel, the presence/absence maps presented in Figure V-8 can be calculated. In this example, a correlation lower than 0.5 was associated with the absence of the studied compound. By using this threshold, API, avicel® and lactose were identified in the 1600 pixels of the image while the low dose compound was only identified in the 6 pixels corresponding to the theoretical simulated data. The threshold was selected to avoid incorrect absences of each compound. As it was shown in Figure V-7, pixels without magnesium stearate provided low correlations because the projected image spectra contained only noise contributions. The differences between low and high correlations were sufficiently obvious to be properly discriminated.

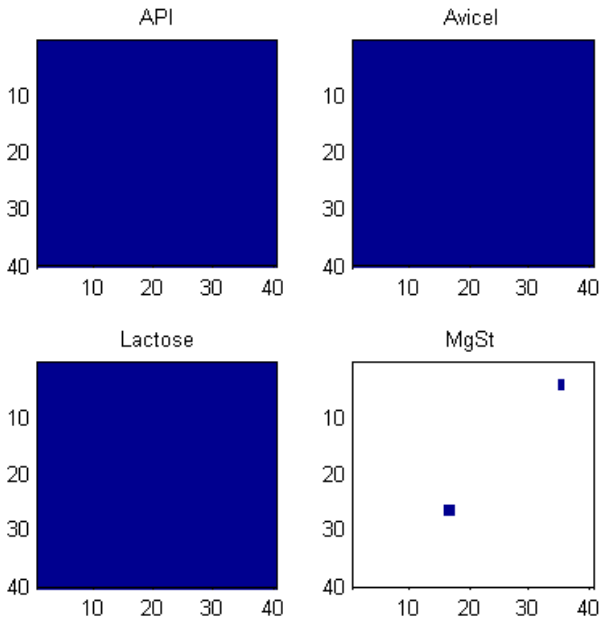


Figure V-8 Presence/Absence maps of compounds in the simulated image (blue colour: Presence of the compound, white colour: absence of the compound)

MCR-ALS was performed with a non-negative constraint on concentrations, with the local rank constraints based on orthogonal projection approach and without PCA-filtering before iterative process. Between MCR-ALS iterations, if a compound was considered to be absent, the concentration was forced to zero as usually done in local rank constraints. In Table V-2, the advantage of using local rank constraints was shown by studying the correlations between the MCR-ALS calculated spectrum and the pure spectrum of magnesium stearate which were respectively equal to 0.78 and 0.89 without or with the local rank constraint.

	MCR-ALS with n-n constraint	**MCR-ALS with n-n and local rank constraints**
$r_{sopt1/API}$	1.00	1.00
$r_{sopt2/Avicel®}$	1.00	1.00
$r_{sopt3/Lactose}$	1.00	1.00
$r_{sopt4/MgSt}$	0.78	0.89
lof %	0.89	0.91
Explained variance %	99.99	99.99

Table V-2 MCR-ALS results on simulated data

4.3. Proposed approach on real dataset

The studied tablet was manufactured with 5 pure compounds (2 actives and 3 excipients). Pure compound Raman images of dimensions 40 pixels per 40 pixels were acquired with the same acquisition parameters as the studied image. A non-centered PCA was applied on each pure compound image. For each compound, the interference basis $\mathbf{K}_{-\mathbf{c}}$ was built with the PCA loadings. The number of loadings for a basis was selected by choosing an appropriate number of components explaining 99.9% of the total variance of each image in order to include the maximum of variability in the interference basis. Projected spectra were studied by correlation with the pure projected spectrum. A conservative threshold of 0.5 was used to avoid incorrect absences. A value lower than the selected threshold indicated the absence of the studied compound. In Figure V-9, the white colour was linked to the absence of the studied compound and the associated concentration was fixed to zero in the MCR-ALS process. However, blue pixels were related to high correlations between a spectrum $\mathbf{r}_{n\perp}$ and the pure projected spectrum $\mathbf{s}_{\perp}$ of a compound *c*.

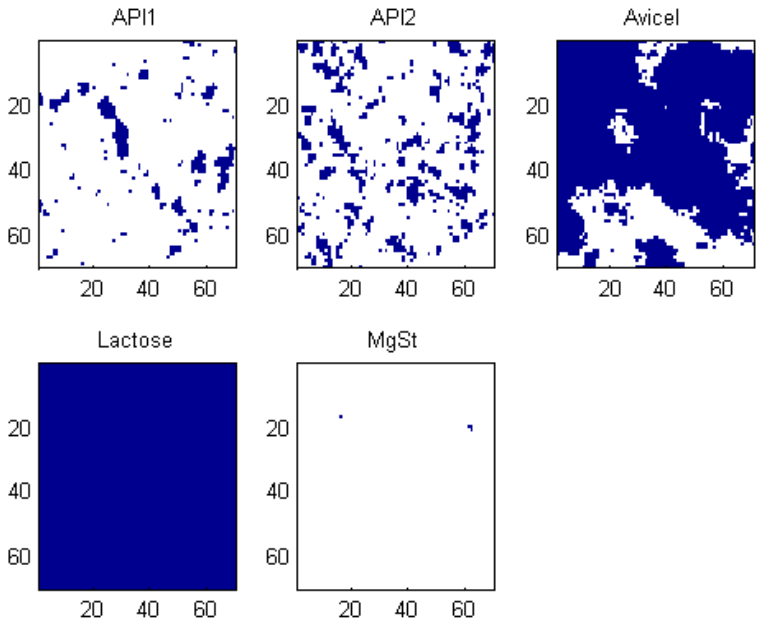


Figure V-9 Presence/absence maps of drug compounds (blue colour: Presence of the compound, white colour: absence of the compound)

Lactose, main excipient of the studied formulation, was detected in the whole surface of the tablet. Absence or presence of the two actives and avicel® were easily highlighted with the

proposed approach. Magnesium stearate, used as a lubricant in the formulation (i.e. the low dose compound) was detected only in few pixels of the image.

MCR-ALS was performed with non-negativity constraint on spectra and concentrations, without using PCA-based filtering. Presence/absence maps were used as local rank constraints between iterations to set up the concentrations to zero if a compound was considered absent. In Table V-3, the importance of the maps of presence/absence matrix in the multivariate curve resolution was highlighted. Indeed, by using the appropriate local rank constraints, better results were obtained. Except for avicel®, correlations between the calculated MCR-ALS spectra and the pure reference spectrum of each constituent were higher by using local rank constraints. Significant improvements were observed for API2 and magnesium stearate by respectively increasing the correlation values from 0.51 to 0.93 and 0.34 to 0.84 (Figure V-10). By using local rank maps constraints, the model was improved as a part of the noise was separated from the model. By reducing noise in the MCR-ALS model, explained variance decreased by taking a value from 99.92% without using local rank constraint to 98.98% with the spatial constraint, and lack of fit increased by taking a value from 2.66% without using local rank constraints to 10.09% with the spatial constraint.

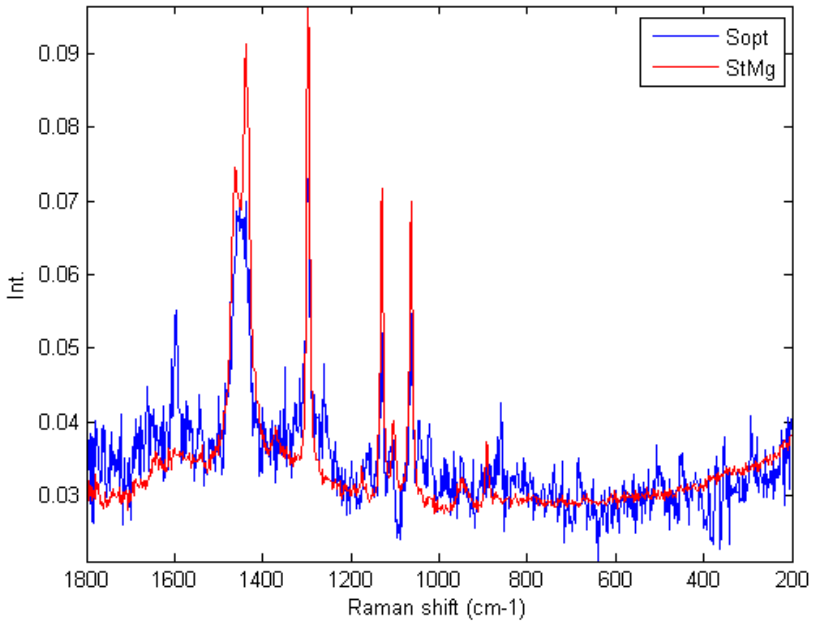


Figure V-10 Calculated spectrum by MCR-ALS (with n-n and local rank constraints) and pure spectrum of magnesium stearate

	MCR-ALS with n-n constraints	MCR-ALS with n-n and local rank constraints
$r_{sopt1/API1}$	0. 70	0. 93
$r_{sopt2/API2}$	0. 52	0. 93
$r_{sopt3/Avicel®}$	0. 95	0. 81
$r_{sopt4/Lactose}$	0. 98	0. 99
$r_{sopt5/MgSt}$	0.34	0.84
lof %	2.66	10.09
Explained variance %	99.92	98.98

Table V-3 MCR-ALS results on real dataset

Resolved distribution maps for all compounds in the formulation were displayed in Figure V-11. Distribution of actives and excipients were as expected according to the nature of the formulation of the drug product. Indeed, as major excipients, lactose and avicel® were identified in the whole tablet surface. The two actives were distributed throughout the tablet. Magnesium stearate, the low dose compound used as lubricant in the formulation, was identified in few pixels of the image.

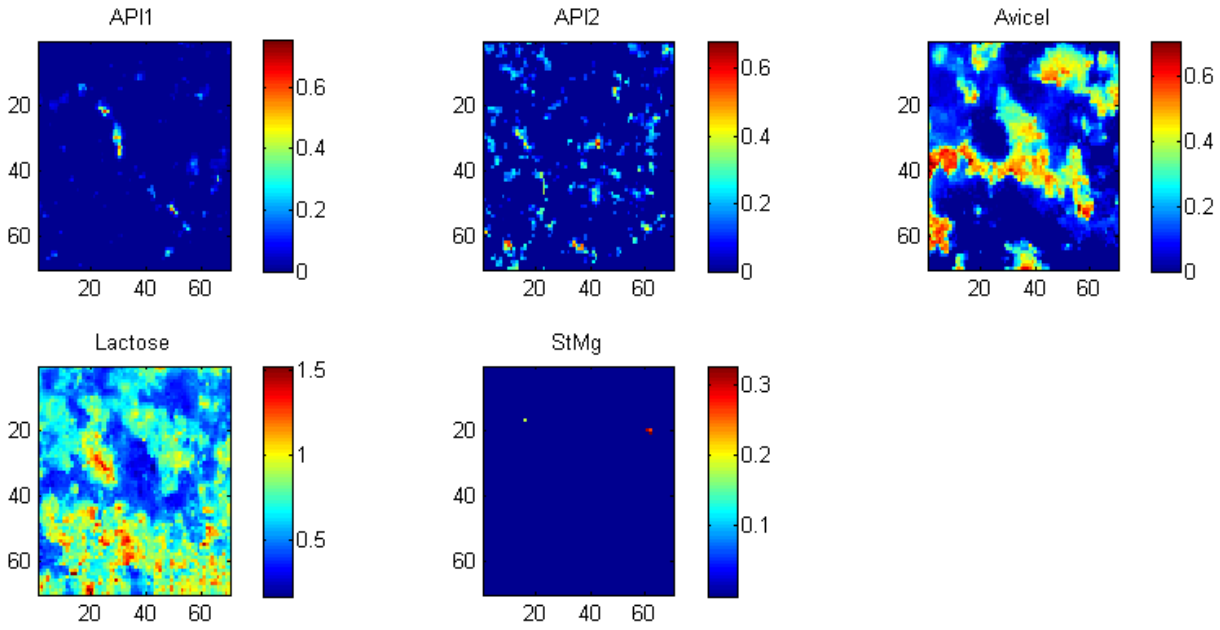


Figure V-11 Distribution maps of the five compounds obtained by MCR-ALS (with n-n and local rank constraints)

5. Conclusions

The proposed approach, which applies orthogonal projections pre-processing to set up presence/absence maps of compounds to be used as local rank constraints in the MCR-ALS iterative process showed excellent results in the case of a low dose compound within a pharmaceutical drug product. For each compound of a pharmaceutical formulation, the spectral matrix (i.e. the unfolded image) was orthogonally projected to a basis containing all the interferences other than the product of interest. Orthogonal projected spectra were then compared with the orthogonal projection of a pure spectrum of interest to the same basis. By choosing a threshold on correlation coefficients between the signals, presence/absence maps can be set up and used as local rank constraint during MCR-ALS process. Results were significantly improved by using this constraint.

Orthogonal projections have been shown to be an interesting approach to set local rank constraints in pharmaceutical formulations because the composition of the sample is known, therefore, interferences spaces can be well defined and because, in case of low dose compounds, the result of algorithms based on variance decomposition may be compromised. In this specific case, orthogonal projection can be viewed as a filtering method which removes detrimental information and focuses on spectral contribution of the product of interest.

Contributions of chapter V

In this work, we proposed an alternative method to set up the absence/presence maps used as equality constraint in MCR-ALS process. This method was based on the hypothesis that each compound has its proper subspace containing spectral variability due to chemical and non-chemical contributions (environmental, acquisition, physical…). This variability can be viewed as a basis of vectors with an appropriate set of dimensions considered as the detrimental subspace. The basis of vectors was calculated for each compound, using non-centered PCA on pure compound images. The number of vectors to be used was selected by explaining 99.9% of the total spectral variance. For each compound, correlations between pure projected spectra and image projected spectra were calculated. A high value indicated the presence of a compound while a low value was associated with the absence of a compound.

By using this approach, maps of absence/presence were in accordance with the simulated and real formulation dataset. This approach appeared as very powerful in the case of a low dose compound because orthogonal projection can be viewed as a filtering method which removes detrimental information and focused on spectral contributions of the product of interest. Moreover, because this method is not based on variance decomposition, it is perfectly well suited for a low dose compound in a mixture dataset. Indeed, using orthogonal projections offered the possibility of working in a different subspace, i.e. the signal subspace. While the usual chemometric tools try to find differences between samples (i.e. by working in the sample space), the proposed approach focused on the signal information. In the case of the low dose compound, information in the sample space is scarce and cannot be extracted by using chemometric methods based on the scatter matrix in the sample space. Using the signal space appeared as a powerful alternative method to circumvent the usual decompositions of chemometric methods.

However, it is important to notice that this approach can only be applied in situations where the space of interferences can be well-defined, i.e. it requires to know the sample composition beforehand. In some cases, during a stability study or when a counterfeit sample is analysed, the studied formulation is not known by the analyst and the proposed approach might be difficult to implement.

This latter observation led us to the following chapter where we will propose an iterative approach, based on a spectral library, spectral distances and orthogonal projections to identify the pure compounds in an unknown pharmaceutical drug product.

Chapter VI: An iterative approach for compound detection in an unknown formulation

1. Introduction 97

2. Materials and methods 97

2.1. Notations 99

2.2. Samples 99

2.3. Raman imaging system 100

2.4. Spectral library 100

2.5. Proposed approach 101

2.5.1. Spectral distances 101

2.5.2. Identification of the pure compound 103

2.5.3. Orthogonal projection 104

2.5.4. Overview of the iterative approach 105

3. Results and discussion 106

3.1. Identification of the tablet compounds 106

3.2. Multivariate curve resolution-alternating least squares 114

4. Conclusions 116

Preamble

In the previous chapter, we highlighted the ability of multivariate curve resolution, associated with the suitable constraint, to provide the distribution of pure compounds in a pharmaceutical drug product, including a low content compound. Most of our previous work was applied on a presupposed known formulation.

However, in some cases, the formulation is not previously known by the analyst and the pure compound identification appeared as a real challenge. Two applications can be cited in the pharmaceutical environment. First application concerns the analysis of a counterfeit sample, which is defined as a product sold under a product name without proper authorization. It may include products without the active ingredient, with an insufficient or excessive quantity of the active ingredient, with the wrong active ingredient, or with fake packaging. Obviously, compounds included in the product are not known beforehand. The second application concerns the chemical modification of a sample during a stability study. During this required test, products are stored in various temperature and humidity conditions, in different packaging, during several months. The objective is to control the product modifications during the storage (for example modifications of the crystalline form or degradations of active).

Because the formulation is not always known, we propose in this section a new methodology to detect pure compound in a mixture dataset. Based on a spectral library, spectral distances and orthogonal projections, this approach should be particularly suited for a formulation which contains a low dose compound.

The proposed approach is tested on a tablet manufactured with one active pharmaceutical ingredient and five excipients, including the low dose lubricant. The tablet is stored 3 months in high temperature conditions before Raman analysis.

Note that this work has been performed with a Raman microscope similar to the one described in the chapters III, IV and V. As this chapter is the reproduction of **Art. IV** submitted in the Journal of Pharmaceutical and Biomedical Analysis in 2015, the readers will find some redundancies between chapters III, IV, V and chapter VI in the materials and methods section.

AN ITERATIVE APPROACH FOR COMPOUND DETECTION IN AN UNKNOWN PHARMACEUTICAL DRUG PRODUCT: APPLICATION ON RAMAN MICROSCOPY[4]

1. Introduction

Raman spectroscopy is becoming increasingly more accepted as a powerful tool in the pharmaceutical research and development environment since this technique has some major benefits [155; 156]. By coupling a microscope with the usual Raman spectroscopy, hyperspectral images providing both spectral and spatial information can be acquired, containing a lot of information on the distribution of active pharmaceutical ingredients (API) or excipients in a product [157]. The development of these analytical methods is very useful to ensure and control the drug product quality during development and beyond post-marketing authorisation [132]. Even if Raman chemical imaging has been used to detect and quantify crystalline forms [158; 159], to characterize particle size [160] or to assess blending effect on tablet quality [114], the main goal in the pharmaceutical industry remains the assessment of the product quality by determining the compound distributions within a tablet [15; 148; 161].

Because of the huge amount of data contained in hyperspectral images, a direct interpretation of the acquired images is not possible and several chemometric tools have previously been published to aid in this task. Hyperspectral data analysis can be divided in several parts depending on the objectives, but most of the times it starts with a pre-processing step followed by a data analysis procedure. Pre-processing methods are usually applied to correct for external perturbations and undesired phenomena to focus on the targeted information. The next step consists of analysing the data by applying qualitative or quantitative chemometric tools such as principal component analysis (PCA) [112; 146], independent component analysis (ICA) [123] or multivariate curve resolution-alternating least squares (MCR-ALS) [84]. These algorithms assume that the acquired spectra are the weighted sum of pure spectra of the formulation compounds. One challenging task during application of these chemometric tools on imaging techniques is how to effectively extract chemical information from the image [48] but, the quality of the extracted signals (related to each pure compound) is also a critical step of the multivariate data analysis.

[4] Mathieu Boiret, Nathalie Gorretta, Yves-Michel Ginot, Jean-Michel Roger. **An iterative approach for compound detection in a unknown pharmaceutical drug product: application on Raman spectroscopy.** Accetpted in *Journal of Pharmaceutical and Biomedical Analysis* (2015).

A lot of algorithms have been previously studied to extract pure spectra (also named endmembers in the remote sensing field) within a mixture dataset [162; 163]. On the one hand, in the chemometrics community, SIMPLISMA (Simple-to-use interactive self-modeling mixture analysis) [87], orthogonal projection approach (OPA) [88], independent component analysis (ICA) [89] or evolving factor analysis (EFA) [90] were used on spectroscopic data to identify these pure signals in a mixture dataset. On the other hand, in the remote sensing community, other approaches such as pixel purity index (PPI) [164], autonomous morphological endmember extraction (AMEE) [165], N-FINDR [166] or vertex component analysis (VCA) [167] appeared as powerful algorithms to extract pure features from hyperspectral images. However, all these algorithms are mainly based on either the hypothesis that each compound has a pure pixel in the image or that a signal contains a sufficient level of spectral contributions for a compound. Considering a low dose compound, it can be assumed that spatial and spectral information is scarce because only few pixels of the image contain the product of interest and because the associated spectral contributions are mixed with the other formulation compounds. Considering this specific case, there are no pure pixels in the studied dataset and identification of the low dose compound appeared as a real challenge [149; 168]. In most chemometric methods, the targeted information is extracted by using the variability between the samples i.e. the differences between the acquired spectra or pixels. But, in the case of the low dose compound, these variations cannot be easily highlighted as the associated contributions are weak and spread into mixture spectra or noise contribution.

In most pharmaceutical applications, the studied formulations are known beforehand, but in some cases, analysts have limited information or do not have prior knowledge on the studied product. For instance, in forensic applications, illegal medicines can be analysed by vibrational spectroscopy to quickly detect counterfeit products [15; 30; 169]. Comparing with genuine drugs, counterfeit samples can be manufactured with different actives or excipients, and identification of product compounds without prior knowledge on the samples could be of interest for analysts. Moreover, during development of a pharmaceutical drug product, stability studies are performed to analyse the evolution of the product in time through different storage conditions (packaging, temperature and relative humidity). Because Raman chemical imaging combined with chemometric algorithms is useful to explore the inner structure of a pharmaceutical drug product [170], evolution of the active quality can be observed in terms of degradations or modifications of its crystalline forms [93]. Therefore, this analytical tool appeared as a very promising methodology to monitor these modifications.

In this article, the objective is the identification of pure compounds in a pharmaceutical tablet, assuming that analysts do not know the studied formulation beforehand and that a potential low dose product is present in the sample. A new methodology is proposed to provide the chemical composition of the tablet. By using a spectral library, compounds are iteratively detected by calculating spectral distances between images and reference spectra. Because each compound has its proper subspace containing chemical information and spectral variability, the associated spectral contributions can be iteratively removed by using orthogonal projections. This approach works exclusively in the signal space, describing the P-dimensional space (one axis per variable) in which the observations can be represented as vectors. Thus, it ensures the detection of a compound without requiring important variations between samples (or pixels). Therefore, by progressively identifying the formulation compounds, from the main product to a product with low contributions, this approach is particularly well adapted to detect all the compounds in a formulation. After spectral identification and in order to provide distribution maps of actives and excipients, MCR-ALS process is applied [81; 82].

The remainder of the paper is organized as follows. Section 2 describes the experimental framework, including notations, samples and apparatus details. The proposed iterative approach will be described in this section. Section 3 presents the ability of the proposed approach to detect the pure compound in an unknown formulation and the MCR-ALS results. Finally, section 4 presents our conclusions.

2. Materials and methods

2.1. Notations

Vectors are noted in bold lowercase, matrices in bold uppercase, and scalars in italic lowercase characters. Vectors are arranged in lines and one line represents one spectrum. The transposed forms of a vector $\mathbf{x}$ and a matrix $\mathbf{X}$ are noted $\mathbf{x^T}$ and $\mathbf{X^T}$, respectively. $\mathbf{I}$ is the identity matrix of dimensions p x p, where p is the number of variables in a spectrum. $\mathbf{x}$ and $\mathbf{X}$ orthogonally projected to a vector basis $\mathbf{K}$ are noted $\mathbf{x}_\perp$ and $\mathbf{X}_\perp$. $\mathbf{\Sigma}$ is the Euclidian orthogonal projector to $\mathbf{K}$.

For a spectral matrix $\mathbf{X} \in \mathbb{R}$, the sample space $\mathbb{R}^n$ describes the N-dimensional space (one axis per observation) in which we can represent the variables (Raman shift) as vectors. The spectral space $\mathbb{R}^p$ describes the P-dimensional space (one axis per variable) in which we can represent the observations (sample spectra) as vectors.

2.2. Samples

A pharmaceutical tablet was especially manufactured by wet granulation for this study. The tablet was prepared by mixing and granulating one active pharmaceutical ingredient, Ivabradine (chronic heart failure treatment), commercialised by "Les Laboratoires Servier", and four excipients: metolose® (Shin Etsu, Tokyo, Japan), eudragit® (Evonik, Essen, Germany), microcrystalline cellulose, and maltodextrin. Dried and calibrated granulates were lubricated with magnesium stearate, which can be associated with a low dose compound as it represented only 0.5% (w/w) of the theoretical formulation. The lubricated granulates were compressed with a rotary press equipped with punches and dies allowing the production of tablets with the required shape. Film-coating and smoothing are carried out in rotative coating pans. The studied drug product contained 10% (w/w) of active in the tablet. The active is known to have several solid state forms (form 1 and form 2) but only the original active form 1 was used to manufacture the product. Submitted to high temperature conditions, the active is known to undergo a crystalline modification from form 1 to form 2. Before Raman chemical imaging analysis, the tablet was stored 3 months at 50°C in a blister. In order to analyse the tablet core and to ensure a flat surface, the tablet was eroded with a Leica EM Rapid system (Leica, Wetzlar, Germany). A visual examination of the tablet did not provide any information concerning the distribution of the different compounds within the tablet.

2.3. Raman imaging system

The tablet image was collected using a RM300 PerkinElmer system (PerkinElmer, Waltham, MA) and the Spectrum Image version 6.1 software. A microscope equipped with an objective 100x magnification was coupled to the spectrometer and spectra were acquired through it with a spatial resolution of 10µm in a Raman diffuse reflection mode. Wavenumber range was 3200–100 cm^{-1} with a resolution of 2 cm^{-1}. Spectra were acquired at a single point on the sample, then the sample was moved and another spectrum was taken. This process was repeated until spectra of points covering the region of interest were obtained. A 785nm laser with a power of 400mW was used. Four scans of three seconds were accumulated for each spectrum. An image of 30 pixels per 30 pixels corresponding to 900 spectra was acquired for a surface of 300µm by 300µm.

2.4. Spectral library

There are many ingredients and actives in the market but this work focused on the products that can be found in SERVIER's formulations. Spectral databases available on the market are mainly

constituted of a single Raman spectrum for each product [171]. In order to include physical and chemical variability of the pure products, a database was built by acquiring one hyperspectral image for each pure compound. The spectral library was constituted of 8 API and 16 excipients, corresponding to 24 images of dimensions 5 pixels per 5 pixels. Images were collected on pure compound tablets using the same conditions as the one described in section 2.3. Details of the spectral library, including uses and functional activities, can be found in Table VI-1.

N°	Product Id.	Actives/Excipients	Main uses and functional category
1	**API 1**	Amlodipine	High blood pressure or chest pain
2	**API 2**	Atorvastatine	Reduces levels of "bad" cholesterol and triglycerides
3	**API 3**	Carvedilol	Heart failure and hypertension
4	**API 4**	Indapamide	Fluid retention (oedema) and hypertension
5	**API 5**	Ivabradine form 1	Heart failure
6	**API 6**	Ivabradine form 2	Heart failure
7	**API 7**	Perindopril	High blood pressure and prevention of heart attack
8	**API 8**	Strontium ranelate	Osteoporosis
9	**Excipient 1**	Aspartame	Sweetener
10	**Excipient 2**	Calcium hydrogen phosphate	Tablet diluent
11	**Excipient 3**	Eudragit RS PO	Sustained release agent
12	**Excipient 4**	Lactose	Tablet diluent
13	**Excipient 5**	Maltodextrin	Binder
14	**Excipient 6**	Magnesium stearate	Lubricant
15	**Excipient 7**	Macrogol	Plasticizer
16	**Excipient 8**	Mannitol	Tablet diluent
17	**Excipient 9**	Microcrystalline cellulose	Tablet diluent
18	**Excipient 10**	metolose	Sustained release agent
19	**Excipient 11**	Povidone	Binder
20	**Excipient 12**	Citric acid	Acidifying agent
21	**Excipient 13**	Starlac	Tablet diluent
22	**Excipient 14**	Sucrose	Sweetener
23	**Excipient 15**	Talc	Anticaking agent
24	**Excipient 16**	Titane dioxide	Opacifier

Table VI-1 – Spectral library

2.5. Proposed approach

In this work, the objective was to identify all the pure compounds of a pharmaceutical drug product, assuming that the chemical composition is not known by the analyst and that a low

dose compound can be present in the studied product. The proposed approach relies on a spectral library, spectral distances and orthogonal projections to iteratively detect pure compounds of a tablet. Since the method is not based on variance decomposition and because it focuses on the signal space rather than the usual sample space, it should be well adapted for a drug product which contains a low dose compound, interpreted as a compound located in few pixels and with low spectral contributions [172].

Let $\mathbf{X}$ be the unfolded image (of dimensions n spectra and p variables) and $\mathbf{R_c}$ be the unfolded image (of dimensions k spectra and p variables) of a reference compound c included in the spectral library. A matrix $\overline{\mathbf{R}}$ is defined by calculating the mean spectrum for each compound c, i.e. the mean spectrum of each matrix $\mathbf{R_c}$. The $\overline{\mathbf{R}}$ spectral matrix contains as many lines as the number of compounds in the spectral library.

2.5.1. Spectral distances

The first step of the proposed approach consists of the spectral distance calculations between every spectrum of the $\mathbf{X}$ matrix and each spectrum of $\overline{\mathbf{R}}$. The spectral angle mapper (SAM) was used for this purpose [173]. SAM is an automated method for directly comparing a signal to a reference spectrum. The algorithm determines the spectral similarity between two spectra by calculating the angle α between the two signals, treating them as vectors in a space with dimensionality equals to the number of bands.

By considering a reference spectrum $\vec{\mathbf{p}}$ and a pixel spectrum $\vec{\mathbf{s}}$ from a two-variable data represented on a two dimensional plots as two points, representation of the α angle is displayed in Figure VI-1. The calculation consists of taking the arccosine of the dot product of the two spectra by applying the following equation:

$$\alpha = \cos^{-1}\left(\frac{\vec{\mathbf{s}}\cdot\vec{\mathbf{p}}}{\|\vec{\mathbf{s}}\|\cdot\|\vec{\mathbf{p}}\|}\right) \qquad \textbf{(VI-1)}$$

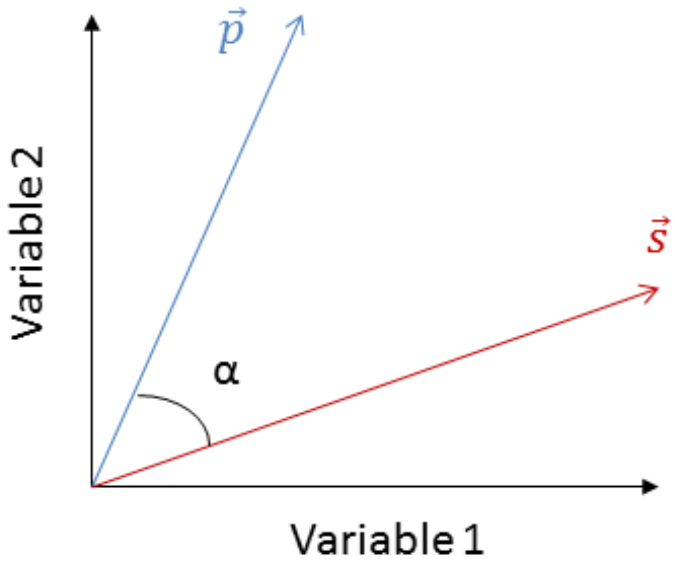


Figure VI-1 – Description of α angle in the spectral angle mapper (SAM) calculation

The SAM algorithm uses only the vector direction and not the vector length, which means that the signal intensity does not modify the angle between two signals. Image spectra and spectral library spectra can thus be acquired with different Raman exposure time, leading to different spectral intensities. Two similar spectra have a SAM value equal to 0 while two orthogonal spectra (dissimilar) have a SAM value equal to $\pi/2$, corresponding to an angle α equal to 90°. Low and high SAM values can be respectively associated with high or low similarity between the signals and the reference spectra.

2.5.2. Identification of the pure compound

Once the SAM values are calculated between every spectrum of $\mathbf{X}$ and each spectrum of $\overline{\mathbf{R}}$, the second step of the proposed approach consists of pure compound identification. SAM values can be observed either by plotting them in function of the pixel number, or by refolding them to get a SAM image, or by using a boxplot representation [174]. This latter representation was considered as the most suitable observation mode since it provides the SAM distributions.

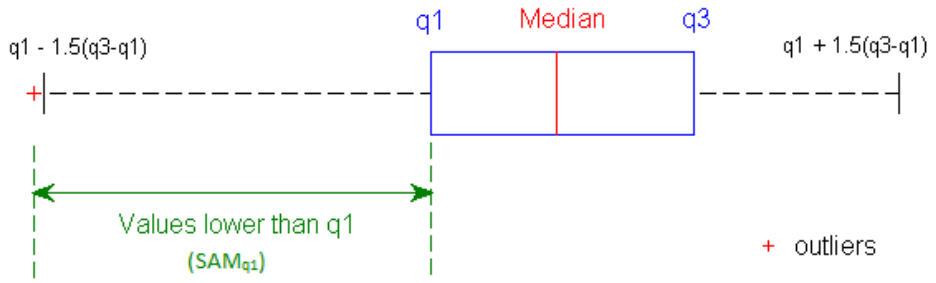


Figure VI-2 – Boxplot representation of SAM

In the proposed approach, mean of the SAM values lower than the quartile q1, including 25% of the sorted values, are observed for each compound of the spectral library (Figure VI-2). This mean value, noted m_{q1} was calculated as follows:

$$m_{q1} = \frac{\sum_{i=1}^{n} SAM_{q1_i}}{n} \qquad \textbf{(VI-2)}$$

Where SAM_{q1_i} are the SAM values lower than $q1$, n the number of samples lower than $q1$. A product which provides the lowest m_{q1} is associated with a pure compound.

A low m_{q1} means that a product is present in most of the pixels in the sampling area. However, in the case of a high m_{q1}, the distribution must be observed. Indeed, in the case of a low dose compound, where only few pixels of the image contain the product, the number of low SAM values is limited, and the m_{q1} will not allow the detection of the product. Boxplot representation offers the possibility to highlight rare SAM values (corresponding to low values in the distribution and considered as outliers) and is thus suitable for the detection of a low dose compound.

Therefore, two criteria (m_{q1} and outliers) have to be observed in order to identify a pure compound in the studied dataset. If for all the compounds of the spectral library, no low values of m_{q1} and outliers (i.e. low SAM values) are observed, then the iterative process stops.

2.5.3. Orthogonal projection

Once a pure compound is identified, the associated mean spectrum is added to the pure spectral matrix **S**. Next, **X** and $\overline{\mathbf{R}}$ are orthogonally projected to a subspace **K** corresponding to the signal

space of the identified compound, including chemical contribution and spectral variability. For a compound *c*, the vector basis is estimated by applying a non-centered singular value decomposition (SVD) on the suitable pure Raman unfolded image $\mathbf{R_c}$. The subspace **K** was built with eigenvectors, selected by choosing a number of components explaining 99.9% of the total variance of each image in order to include the maximum of spectral information (including spectral variability) in the basis.

An orthogonal projector **Σ** to **K** is calculated as follow:

$$\mathbf{\Sigma = I - K^T(KK^T)^{-1}K} \qquad \textbf{(VI-3)}$$

By orthogonally projecting the spectra **X** and mean pure spectra $\overline{\mathbf{R}}$ to the basis **K**, projected image spectra $\mathbf{X}_\perp$ and pure projected spectra $\overline{\mathbf{R}}_\perp$ are obtained by:

$$\mathbf{X}_\perp = \mathbf{X\Sigma} \qquad \textbf{(VI-4)}$$

and

$$\overline{\mathbf{R}}_\perp = \overline{\mathbf{R}}\mathbf{\Sigma} \qquad \textbf{(VI-5)}$$

The following iterations are performed on $\mathbf{X}_\perp$ and $\overline{\mathbf{R}}_\perp$.

2.5.4. Overview of the iterative approach

The proposed approach is thus divided with the different steps listed below and graphically represented in Figure VI-3:

1. Initialise the calculation process with $\mathbf{X}_\perp = \mathbf{X}$ and $\overline{\mathbf{R}}_\perp = \overline{\mathbf{R}}$
2. Calculate SAM values between every spectrum $\mathbf{X}_\perp$ and each mean spectrum of $\overline{\mathbf{R}}_\perp$ from the spectral library
3. By observing m_{q1} and outliers, identify a pure compound *c* within the formulation and add the corresponding mean spectrum to **S**. Iterative process stops if the two observed criteria do not provide low SAM values, corresponding to spectral similarity between the signals

4. Spectra **X** and $\overline{\mathbf{R}}$ are orthogonally projected to the subspace **K** constituted of the non-centered SVD eigenvectors calculated on unfolded pure Raman images of the identified compounds

5. Back to step 2

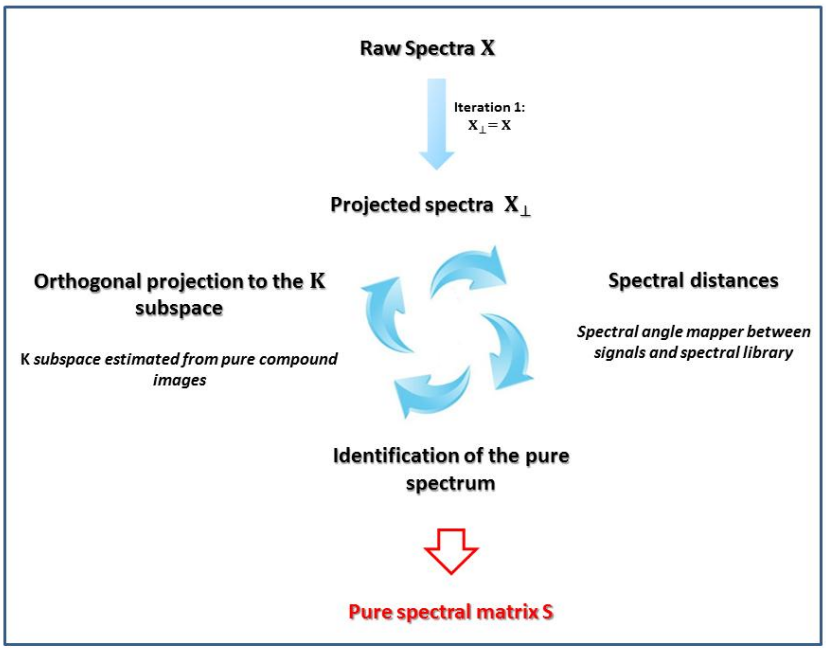


Figure VI-3 – Description of the proposed approach

3. Results and discussion

3.1. Identification of the tablet compounds

First, Raman hyperspectral dataset was spike-corrected to reduce the effect of cosmic rays [61]. Next, the 3-dimensions data cube was unfolded to obtain a 2-dimensions matrix **X** of dimensions *n* samples (spectra) per *p* variables (Raman shift). The Spectral range was reduced in order to focus only in the region of interest corresponding to Raman shift from 1800 to 200 cm^{-1} (Figure VI-4).

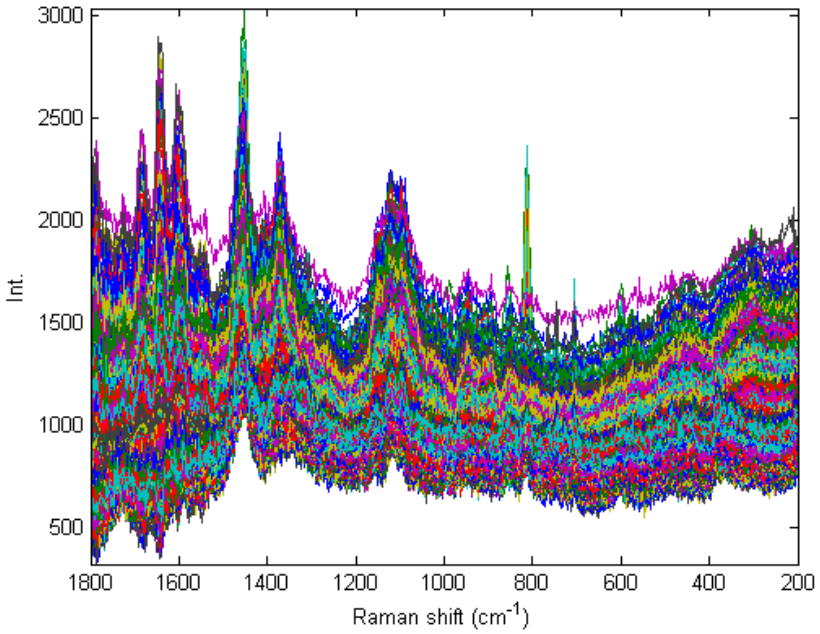


Figure VI-4 – Raw spectra of the image of dimensions 30 pixels per 30 pixels (900 spectra)

In Figure VI-5, the 24 mean reference spectra of $\overline{\mathbf{R}}$ were displayed. Note that for several products, the correlation between spectra can be important. For example, correlations between Ivrabradine form 1 and form 2 or between lactose and starlac® were respectively equal to 0.75 and 0.90.

Image was acquired with a spatial resolution of 10µm, corresponding to a pixel size higher than the particle sizes of the tablet compounds. Therefore, it can be assumed that each pixel of the image contained a mixture of the formulation compounds. Even if pure pixel can be present for the main formulation excipients, identification of all the tablet compounds, including low dose products, appeared as a real challenge.

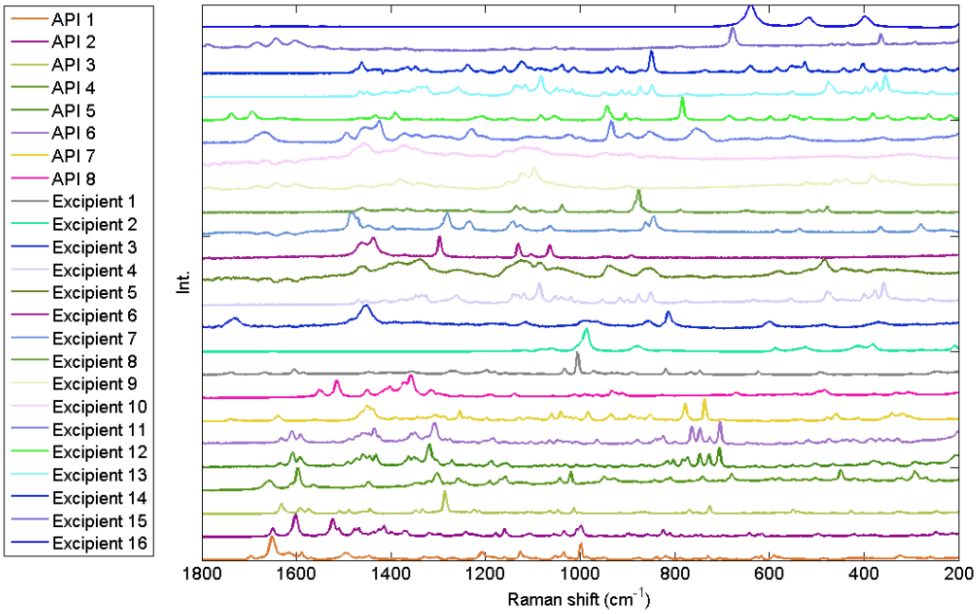


Figure VI-5 – Mean spectra of the 24 pure products included in the spectral library

Iterative process was applied on the pre-processed Raman spectral matrix **X**. During the first iteration step, SAM values were calculated between the 900 pre-processed Raman spectra and the 24 mean reference spectra obtained from the unfolded pure images. Results of the first step of the iterative process are displayed in Figure VI-6.

In iteration 1, the lowest mean of SAM values lower than *q1* (m_{q1}), equal to 0.07, was calculated for metolose®, the main excipient of the studied formulation. Therefore, this compound was identified as a formulation compound and the associated mean spectrum was added to the matrix **S**. In the next step, the original pre-processed matrix **X** was orthogonally projected to a subspace **K**, built by applying a non-centered SVD on the pure unfolded image of metolose®. This basis of vectors includes the spectral contributions of the identified compound. By orthogonally projecting all the spectra to this basis, metolose® spectral contributions were subtracted and only the information of the other compounds was kept. The number of eigenvectors was chosen by explaining 99.9% of the total spectral variance, and was equal to 5. **X** and $\overline{\mathbf{R}}$ spectra were orthogonally projected to the subspace **K**, providing new $\mathbf{X}_{\perp}$ and $\overline{\mathbf{R}}_{\perp}$ matrices.

The second iteration calculated SAM values between every orthogonal projected spectrum $\mathbf{X}_{\perp}$ and each projected mean spectrum of $\overline{\mathbf{R}}_{\perp}$. By definition, all the spectral contribution and

variability (due to the environment, acquisition or physical variations) from metolose® were subtracted. The lowest m_{q1}, equal to 0.57, was calculated for eudragit®, the second main excipient in the studied formulation. The associated mean spectrum was added to the pure spectral matrix **S**. Initial matrices **X** and $\overline{\mathbf{R}}$ were orthogonally projected to the basis **K** including vector basis linked to metolose® and eudragit®. The number of eigenvectors used to build eudragit® subspace was equal to 5.

In iteration 3, crystalline form 1, crystalline form 2 and microcrystalline cellulose provided three low values of m_{q1} respectively equal to 0.92, 0.87 and 1.00. The closeness of m_{q1} values between the two crystalline forms of active can be explained by the correlation between the two pure spectra, equal to 0.75. Indeed, because of high spectral correlation between those two pure spectra, SAM values between the projected spectra and the pure spectra were close. Since only slight differences between m_{q1} values were observed, the selection of a compound among these three pure products can be discussed. In this work, the minimum m_{q1} value was used to identify a pure compound and hence the identified compound of the third iteration was the second form of API. This crystalline form of active was not used in the initial manufacturing process and appeared during the 3 months of storage at 50°C in a blister. The corresponding spectral contribution, calculated by using 4 eigenvectors, was added to the **K** subspace. Note that the minimum SAM values could have been used to detect a pure compound, and would have been led to a different order to extract the pure products. Using orthogonal projections offers the possibility of working in the signal space $\mathbb{R}^{\mathrm{p}}$ rather than in a sample space $\mathbb{R}^{\mathrm{n}}$. By applying this methodology, the spectral information linked to a pure compound can be iteratively identify and subtracted. Thus, the extraction order of a compound is not a critical parameter of the proposed approach.

By observing the boxplot representations in Figure VI-6, several outliers (lower than m_{q1} and represented as red crosses) were highlighted for several compounds (Spectral number 5, 13 and 14 in iteration 5 for example). Outliers can be associated with low SAM values between the pure projected spectra and the image projected spectra and thus to heterogeneous distribution of a product in the tablet. Indeed, broad SAM value distributions including outliers (i.e. low SAM values) rely on compounds which are not homogeneously distributed in the sample.

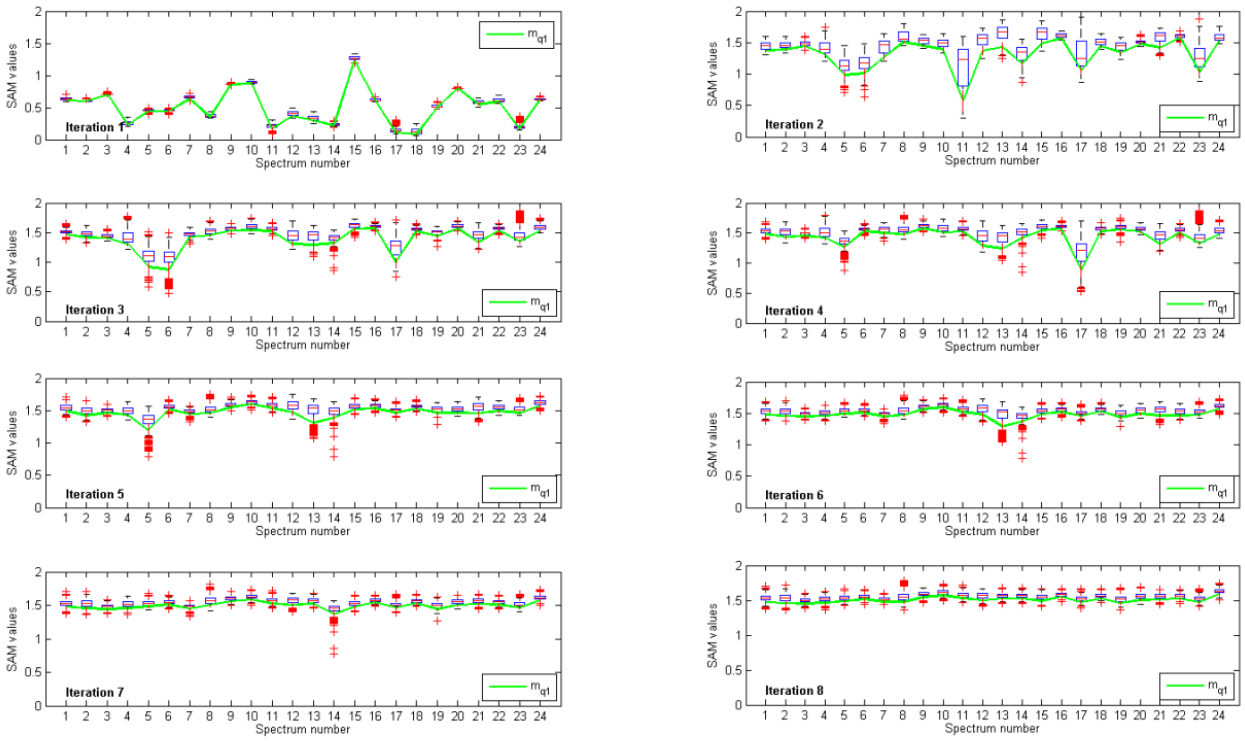


Figure VI-6 – SAM values boxplot and m_{q1} values calculated from iteration 1 to 8 (to identify each sample number, readers are refered to Table VI-1)

In iteration 4, as a consequence of the identification of the second crystalline form of active in iteration 3, and because of the spectral correlation between the two crystalline forms of active, the API form 1 m_{q1} value significantly increases, taking a value from 0.92 to 1.26. The lowest m_{q1} value, equal to 0.87, was calculated for microcrystalline cellulose. The corresponding spectral contribution, calculated by using 5 eigenvectors, was added to the **K** subspace.

In iteration 5, the lowest m_{q1} value equal to 1.19, was calculated for the first crystalline form of API. By orthogonally projecting the original **X** and $\overline{\mathbf{R}}$ matrices to the subspaces of the five first identified pure compounds, calculations focused on compound which provides lower contributions, due to a low dose compound or because of a low Raman response. On the one hand, in the case of a compound present in few pixels, only few low SAM values are calculated, and thus the m_{q1} values cannot be sufficient to identify a compound. One the other hand, in the case of a low spectral response, the signal can be mixed with noise contribution and SAM values can be higher than the values calculated in the previous iterations. In both cases, in addition to m_{q1} values, observations of outliers can be useful.

The sixth iteration highlighted the lowest m_{q1} for matltodextrin, with a value equal to 1.29 and several outliers identified below m_{q1}. Despite of a concentration of 6.5% in the theoretical formulation, identification of the maltodextrin in the sixth iteration and the high m_{q1} value can be explained by a weak Raman response of this product comparing with the other formulation compounds.

The seventh iteration highlighted high m_{q1} values for all the compound, except for the magnesium stearate which provided a m_{q1} value equal to 1.36. Moreover, by observing the SAM value distribution, several pixels with low SAM values can be highlighted, taking values from 0.77 to 1.30. SAM values calculated between magnesium stearate projected pure spectrum and the associated projected spectra $\overline{\mathbf{R}}_{\mathbf{MgSt}_\perp}$ are displayed in Figure VI-7. Only three SAM values lower than 0.8 were calculated. By observing the associated projected spectrum $\mathbf{X}_\perp$ (at a specified position y = 23 and x = 4) and the projected mean spectrum $\overline{\mathbf{R}}_{\mathbf{MgSt}_\perp}$, a correlation equal to 0.75 between the two signals can be highlighted (Figure VI-8).

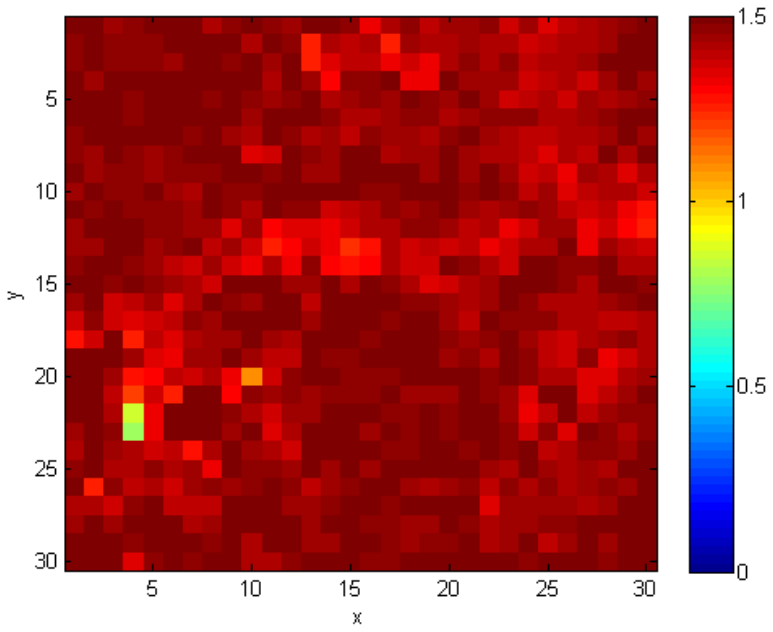


Figure VI-7 – SAM values between pure projected spectrum $\overline{\mathbf{R}}_{\mathbf{MgSt}_\perp}$ of magnesium stearate and the $\mathbf{X}_\perp$ matrix (iteration 7)

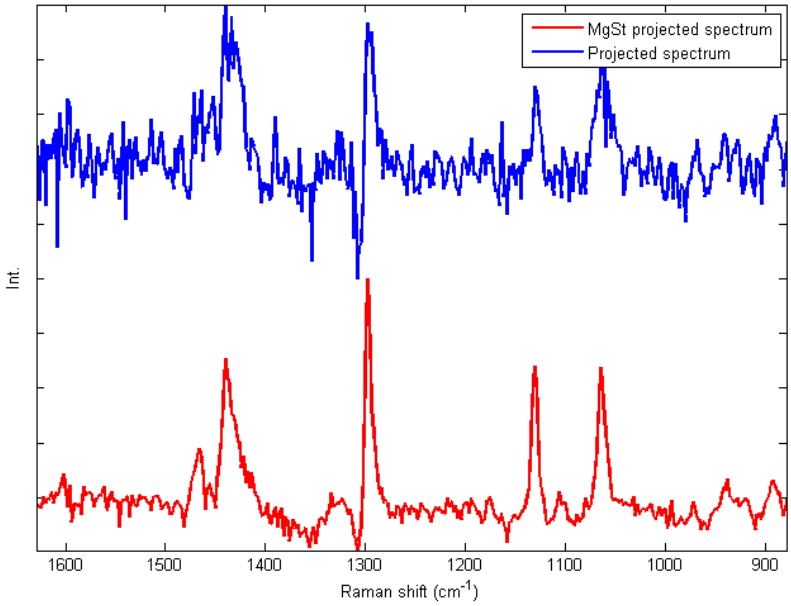


Figure VI-8 – Projected spectrum $\mathbf{X}_{\perp}$ at positions y = 23 and x = 4 (blue) and the projected mean spectrum $\bar{\mathbf{R}}_{\mathbf{MgSt}_{\perp}}$ (red)

Despite of the low magnesium stearate concentration in the formulation (0.5% w/w), the product was successfully detected by the proposed approach. The subspace **K** was completed with the magnesium stearate subspace built by using 3 eigenvectors and **X** and $\bar{\mathbf{R}}$ were orthogonally projected to this new vector basis.

Iteration 8 provided high m_{q1} values for all compounds of the spectral library. Moreover, no outliers, corresponding to spectra with low SAM values were identified. No additional compounds were identified and the iterative process stopped.

At the end of the iterative process, the pure spectral matrix **S** was constituted of seven spectra corresponding to the following compounds: metolose®, eudragit®, API form 2, microcrystalline cellulose, API form 1, maltodextrin and magnesium stearate. Assuming that the formulation was unknown, the proposed approach successfully identified the entire tablet composition, including the modification of a crystalline form and the low dose compound.

In this work, the compound selection and the end of the iterative process were mainly based on the observations of m_{q1} and SAM outlier values. On the one hand, a compound highly concentrated and homogeneously distributed in a sample provides low m_{q1} values and can be easily identified. On the other hand, a low dose product, distributed in a few pixels with a low

spectral response, provides only few low SAM values related to outliers. Observations of m_{q1} and outliers during the iterative process offers the possibility to detect both major and minor compounds in the sample. Since the approach works exclusively in the signal space, it ensures the detection of a compound without requiring important variations between samples. Therefore, by progressively identifying the formulation compounds, from the main product to a product with low contributions, it is suitable to detect all the compounds in a formulation

Even if the method appeared as a powerful methodology, it requires a minimum of expertise, especially to interpret the SAM distributions (m_{q1} and outliers). A semi-automatic method was tested in order to select the number of compounds (i.e. the number of iterations) to be used. In theory, once all the compounds are identified, all m_{q1} values should tend to a value close to $\pi/2$ (α equal to 90° due to the dissimilarity between signals) and the standard deviation between all the calculated m_{q1} values should reach a plateau. Figure VI-9 displays the evolution of standard deviation calculated on m_{q1} for 15 iterations. The standard deviation progressively decreased from the first to the seventh iteration, taking a value from 0.25 to 0.05. After the seventh iteration, the standard deviation reached a plateau, meaning that no additional compound was present in the formulation. In addition to the visual interpretation of m_{q1} and outliers for each compound of the spectral library, this approach could be useful to select the number of compounds in the studied sample.

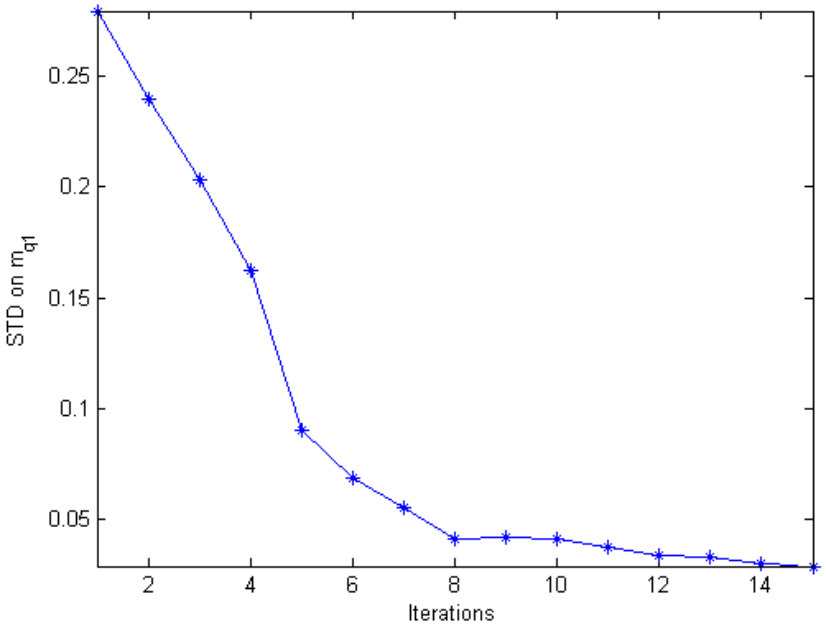


Figure VI-9 – Evolution of standard deviation of the 24 m_{q1} values

The proposed iterative approach appeared as an interesting methodology to detect the compounds of a pharmaceutical drug product without prior knowledge on the formulation. Using these identified spectra and in order to provide distribution maps of each compound, multivariate curve resolution-alternating least squares (MCR-ALS) was applied on the data.

3.2. Multivariate curve resolution-alternating least squares

MCR-ALS was performed with non-negativity constraint on spectra and concentrations, without using PCA-based filtering in order to keep the maximum of information before the iterative process [149] and by using the initial estimate spectral matrix **S** determined in the previous section. To enhance slight spectral variations, data were pre-processed with a Savitzky-Golay [59] first derivative with a 2nd order polynomial smoothing on a 9 points window. Model was considered as optimum after 6 iterations, providing a lack of fit equal to 5.9% and a percentage of variance explained equal to 99.9%.

The distribution of actives and excipients is displayed in Figure VI-10. The main excipients, metolose®, eudragit® and microcrystalline cellulose were easily detected on the tablet surface. Distribution of actives highlighted the transformation of active form 1 to form 2 during the stability study. An estimation of the compound concentration was calculated by applying the following method for each of the *c* product:

$$\mathbf{C_c} = \frac{\sum_{i=1}^{n} \mathbf{c_i}}{\sum_{j=1}^{c} \sum_{i=1}^{n} \mathbf{c_{i,j}}} \times \mathbf{100} \qquad \textbf{(VI - 6)}$$

where *n* is the number of spectra (or pixels).

Even if these values have only an indicative meaning, since only a small area of the mixture was considered, the results were in accordance with the theoretical tablet formulation (Table VI-2).

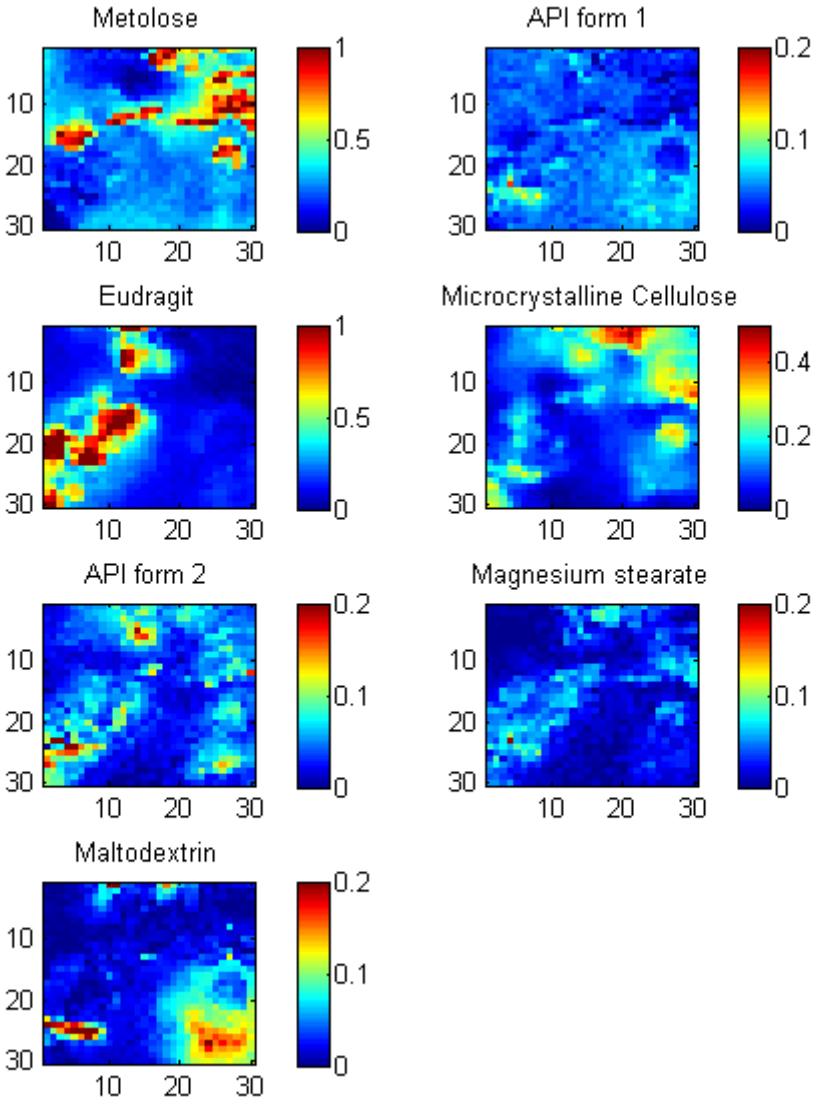


Figure VI-10 – Distribution maps of metolose®, API form 1, eudragit, microcrystalline cellulose, API form 2, magnesium stearate and maltodextrin

Pure compound	Theoretical amount (% w/w)	Calculated C_c (% w/w)
Metolose®	40	39
Eudragit®	25	29
API	11	5 (form 1) + 5 (form 2)
Microcrystalline cellulose	17	15
Maltodextrin	6.5	6
Magnesium stearate	0.5	1

Table VI-2 – Estimation of the compound concentration

4. Conclusions

In this article, a new methodology to identify pure compounds in a pharmaceutical drug product by Raman microscopy was proposed. It was assumed that the chemical composition of the product was not known beforehand by the analyst and that a low dose compound can be present in the studied tablet. In this proposed approach, a spectral library, spectral distances and orthogonal projections were used to iteratively detect pure compounds of a tablet. Since the method is not based on variance decomposition, it was well adapted for a drug product which contains a low dose compound, interpreted as a compound located in few pixels and with a low spectral response.

The method was tested on a tablet specifically manufactured for this study and constituted of one active pharmaceutical ingredient and 5 excipients, stored 3 months at 50° in a blister before analysis. A spectral library, constituted of 8 actives and 16 excipients, was used as a spectral database. Two forms of the active pharmaceutical ingredient were detected. A modification of the crystalline form during the storage was highlighted. Moreover, the lubricant, considered as a low dose compound, was successfully identified in the tablet. By using the pure identified spectra, multivariate curve resolution-alternating least squares was applied on the whole spectral matrix. Results provided the distribution maps of each compound in the tablet. The distribution and estimation of the amount of each compound was in accordance with the theoretical formulation.

This approach could be particularly interesting for analyst in either the case of identification of pure compound in an unknown product such as a counterfeit product or during the stability study of a pharmaceutical drug product (tablets, powders or extrudates) to study the degradation of active and the modification of crystalline forms.

Contributions of chapter VI

In this chapter, we proposed an iterative method to identify pure compounds of a pharmaceutical drug product, assuming that the chemical composition is not known beforehand and that a low dose compound can be present in the studied formulation.

The proposed approach requires a spectral library constituted of pure spectral images of pure compounds which are subject to be present in the formulation. Obviously, the more the number of compounds in the spectral library, the more efficient the proposed approach. Spectral distances were calculated between every image spectra and reference mean spectra of the spectral library. A low distance value was associated with the identification of a pure compound. Image spectra and pure spectra were then orthogonally projected to the vector subspace of the identified compound, including its chemical and physical variability. The corresponding information is subtracted from the data and the next iteration can be performed. The process was repeated until that no pure compounds were identified (high values of spectral distances). Due to low spectral resolution (lower than the particle sizes of the different constituents) and to the penetration depth of the Raman signal, a pixel cannot be assigned to a pure spectrum. Thus, as presented in chapter V, the use of the signal space (by using orthogonal projections) appeared as particularly well-adapted to iteratively subtract the contribution of a selected pure compound. Indeed, since a loss of information can occur by applying variance decomposition, or more generally by applying algorithms based on the decomposition of a statistical moment, the use of orthogonal projections offers the possibility to work in a signal space, which ensures the conservation of the low dose contribution between iteration.

In this work, a tablet stored 3 months in a blister at 50°C was studied. Once the pure products were identified in the hyperspectral dataset, MCR-ALS was performed to provide distribution maps of each compound. Two conclusions can be highlighted. First, during the storage, we noticed a crystalline modification of the active, due to high temperature conditions. Although that the two crystalline form spectra were highly correlated, the proposed approach successfully detects the apparition of the second form. Second, the proposed approach was able to detect the low dose lubricant and the maltodextrin, which is known to provide weak Raman response.

Chapter VII: Conclusions and future work

1. Introduction 119

2. Main contributions 119

2.1. A flashback to the beginning of this work 119

2.2. Applications of a blind source separation methodology 121

2.3. Application of multivariate curve resolution 122

2.4. An alternative for presence/absence map estimations 123

2.5. Compound detection in an unknown formulation 123

3. Limits and future work 124

1. Introduction

In this work, we aim at challenging chemometric tools on Raman microscopy for studying both identification and distribution of compounds in a pharmaceutical drug product. The common thread running through the current thesis is focused on the identification of a low dose compound within hyperspectral images. Although apparatus or experimental parameters can significantly improve the results, this work is only focused on the multivariate data analysis aspect to extract the targeted information.

In a pharmaceutical drug product, but more generally in a mixture dataset, a low dose compound can be present. In this work, it is defined as a product distributed in a few pixels with a low spectral response (comparing with the other compounds) leading to weak variance in the acquired hyperspectral dataset. As usual chemometric tools are mainly based on variance decomposition, requiring a sufficient level of contributions in the spectral dataset, the detection of a low dose compound and the assessment of its distribution in the sample appeared as a real challenge.

In this thesis, we tested the ability of ICA and MCR-ALS to reach the objectives previously described. Regarding the case of low dose compound identification, we proposed new methodologies of working for applying MCR-ALS and for estimating the presence/absence maps used as equality constraint during optimisation procedure. Finally, we proposed an iterative method to detect pure compounds in an unknown formulation, assuming that the studied product is not known. This final chapter summarizes contributions of the thesis by emphasizing key points of the developed chapters. Scientific perspectives and future works will be presented and discussed.

Although only pharmaceutical samples and Raman microscopy are studied in the current manuscript, conclusions and proposed approaches can be extended to a general case of hyperspectral datasets which potentially includes a low dose compound (example: detection of a contaminant in food engineering).

2. Main contributions

2.1. A flashback to the beginning of this work

Hyperspectral datasets provide both spatial and spectral information from a sample. The spatial dimension provides distribution maps of the studied compounds while the spectral dimension

can be associated with chemical information. Hyperspectral chemical imaging is particularly appreciated in different fields and it is now considered as a powerful analytical tool in the pharmaceutical environment. As it was listed in this current thesis, a lot of applications have been previously developed, to study the distribution of actives and ingredients for instance, but, none of them focused on a low dose compound spread into hyperspectral dataset (see Chapter II, paragraph 3). Thus, we decided to focus on this specific case where a low dose product is distributed in a mixture dataset.

Two objectives were pursued, based on Raman hyperspectral images:

i/ Study of the compound distributions within a pharmaceutical drug product

ii/ Identification of a low dose compound in a pharmaceutical drug product

Across the current thesis, the first objective was dealt with the help of ICA and MCR-ALS. These chemometric methods were applied to study the compound distributions in hyperspectral images. Methods were challenged and discussed on a real case example of Raman images of pharmaceutical tablets. Because these chemometric tools mainly use the decomposition of statistical moments or apply filtering process to reduce the matrix dimensions before calculations, the difficulty of extracting the information linked to a low dose compound was expected. Limitations of these approaches to reach our objective were rapidly verified and new ways of working were proposed.

To our point of view, the main issue to solve the second objective (i.e. the low dose compound identification) can be summarised with one specific well-known mathematical parameter: the variance. This latter is defined as one of the moments of a distribution. It describes how far a set of samples is spread out around the mean. In this work, because the data are not centered, the variance can be associated with the dispersion of samples around a predefined reference.

When a low dose compound is present in hyperspectral images, low spatial and spectral contributions provide weak variance in the entire dataset. Because most usual algorithms are based on variance decomposition, we had to think about new approaches in order to extract the targeted information with multivariate data analysis. Note that the useful information can be defined by both the low dose compound contribution and the other compound contributions in the dataset.

2.2. Applications of a blind source separation methodology

First, from a practical point of view, ICA is known to be fully suited to provide an estimation of pure signals from a mixture dataset. When the number of independent components is well-chosen, calculated signals are highly correlated to pure spectra. By projecting the initial matrix on the calculated signals, it is very easy to display distribution maps of each compound, provided that spectra have not similar profiles.

In this work, the JADE algorithm was chosen to perform the calculation. Even if the determination of independent signals is not based on variance decomposition, it requires high order statistics (i.e. fourth order cumulant) to set up the independent signals. Before independent signal determination, whitening and reducing steps are performed in order to reduce the size of the data. By applying these reducing steps, a loss of information, especially for a low dose compound, can occur.

The potential of blind source separation method to extract pure signal information from an unknown dataset was verified in this thesis. In addition, we highlighted the importance of choosing the suitable number of components to develop a model. Moreover, we presented some results on the detection of a low dose compound (signal and distribution). By calculating extracted signals with a reference pure spectrum of the low dose compound, the product was successfully identified. But, the model had to be built with a number of components higher than either the theoretical number of compounds in the formulation or than the physico-chemical rank of the mixture dataset. Therefore, it appeared as difficult, if not impossible, to identify a low dose compound if the reference spectrum of the product is unknown. In addition, using this over-decomposed ICA model significantly decreased the quality of signals and distribution maps for the main actives and excipients. Thus, depending on the objective, the proper number of components must be selected. In addition, we highlighted that distribution maps have to be interpreted with carefulness since there are estimated by a projection of the original matrix on the calculated signals. In the case of similarity between pure compound spectra, distribution maps can be incorrect.

In conclusion, to extract the low dose compound information, by using ICA and especially the JADE algorithm (which includes a filtering step before starting the independent signal determination), it is important to use a sufficient number of dimensions in order to keep the targeted information in the matrix. However, using an over-segmented model significantly decreases the signal quality for other formulation compounds. A compromise must be properly selected, depending on the objectives.

2.3. Applications of multivariate curve resolution

Second part of this thesis mainly focused on MCR-ALS to study the distribution of actives and excipients in the studied pharmaceutical drug product. As previously mentioned, the main objective was the detection of a low dose compound in a pharmaceutical formulation, assuming that it is present in a few pixels, with low spectral contributions. A new methodology was proposed, ensuring that the information from the low dose compound is maintained in the spectral matrix before starting the iterative process of alternating least squares.

Three different cases were challenged, especially by modifying the matrix before alternating least squares iterations. The first case was considered as the usual way of working, applying a PCA-filtering on the spectral matrix using a number of components equal to the number of compounds in the formulation before starting MCR-ALS. The second also used a PCA-filtered matrix, but the number of components was progressively increased from k (i.e. the number of compounds in the formulation) to the maximum of variables (i.e. the data corresponds to the non-filtered matrix). The third case used an augmented matrix where the low dose compound spectral information was added to the initial dataset.

We concluded that in the case of a low dose compound, it is very important to ensure that the corresponding information is kept in the initial matrix before starting the iterative process. Therefore, two approaches should be considered in the case of a low dose compound: i/ a PCA-filtering step should be avoided or ii/ the low dose spectral information should be added to the initial matrix (i.e. augmented matrix).

In addition, we emphasized the necessity of using both proper pre-processing tools on the data and constraints on concentrations and spectra during the MCR-ALS optimisation procedure. The latter improves significantly the results of the calculation but might be difficult to set up in this particular case. Alternative method for setting up this constraint was proposed in Chapter V.

Chapter III and chapter IV highlighted the difficulty to extract a low dose compound contribution, either by using ICA or MCR-ALS. The only way of extracting the targeted information is to limit the filtering process usually applied before setting up the model. This filtering process is mainly based on the decomposition of statistical moments and could lead to the loss of contributions in the case of a low dose compound. In order to circumvent these limitations, another paradigm, based on the signal space rather than the sample space, was proposed and tested in chapter V and chapter VI.

2.4. Alternative method for presence/absence map estimations

Constraints used in alternating least squares optimisation procedure are essential to reduce intensity or rotational ambiguities and to move towards a unique solution. Different constraints have been previously studied and applied but one of them, called the equality constraint, is known to significantly improve the resolution.

Usually, equality constraint is based on local rank maps to estimate the compound absence/presence maps. A well-known and famous method, called FSMW-EFA [144], has been previously used on Raman chemical imaging. This method is based on the singular value decomposition of a pixel and its neighbourhood. However, in the case of a low dose compound which has low spectral contribution, the spectral information can be mixed with the other compound contributions or spread into the noise. Thus, a method based on singular value decomposition and correlations between spectra could fail to identify this product.

In this work, the objective was to provide an alternative to the presence/absence map determination by circumventing the variance limitation linked to a low dose compound. For each product of the mixture, the proposed approach used orthogonal projection to a space containing the contributions to be removed (interference or detrimental subspace), i.e. information from compounds other than the compound of interest. The projected spectra were analysed and presence or absence of a compound were highlighted by using correlation maps. We tested and validated this approach on a simulated dataset which was manufactured with a low content product in 6 pixels, and then on a real dataset.

By using only the spectral information, the proposed approach focuses on the signal space. Therefore, the limitations encountered by working in a sample space were circumvented. Indeed, since this methodology is only based on spectra, it does not require significant variations between samples. A low dose compound, which was defined as a product with low spatial and spectral contributions, can then be identified. Note that this approach can only be applied when the sample composition is known because the space of interferences must be well-defined.

2.5. Compound detection in an unknown formulation

Previous developments in the thesis assume that the studied formulation is known. However, in some applications such as counterfeit detection or analysis of a product during a stability study, it could be useful to identify the different compounds, associated with their distributions in the sample, without knowing the formulation.

We proposed a methodology to identify all the compounds of a pharmaceutical formulation, including a potential low dose compound, assuming that the analyst do not know the formulation beforehand. This approach uses a spectral library, which includes hyperspectral images of pure compounds, spectral distances, and orthogonal projections. Spectra are iteratively detected in the studied sample by observing and interpreting the spectral distance values. It is important to have a spectral library which includes a large number of actives and excipients. The more the number of compounds in the spectral library, the more efficient the proposed approach to detect an unknown compound in a formulation.

Based on orthogonal projections, this method is suitable for a formulation which includes a low dose compound. Indeed, it is only based on the spectral information and does not require high differences between samples. By using the signal space rather than the sample space, this method does not require the calculation of statistical moments between samples, then, even if the sample is scarce and scattered in the image spectral mixture, it will be identified with success.

Once the pure compound spectra detected, curve resolution methods can be applied to provide the distribution of actives and excipients in the pharmaceutical drug product. The proposed approach was tested on a formulation which contains different forms of actives and excipients, including a low dose lubricant.

3. Limits and future work

The work in this thesis presents some limitations which will be discussed in this section. Across the thesis, ICA and MCR-ALS were challenged on Raman images of a pharmaceutical drug product to detect major and minor compounds in the formulation. Limitations of these algorithms were rapidly highlighted by focusing on the case of a low dose compound and the proposed approaches significantly improve the ability of algorithms for low dose compound detection.

Although ICA was successfully applied on the data to detect pure signals of the main actives and excipients (chapter III), it is important to keep in mind that the JADE algorithm, used in this work, starts with a reduction of dimensions based on singular value decomposition. In the case of a low dose compound, the spectral variance is weak in the matrix and the reduction of dimensions can lead to a loss of the associated information. In order to circumvent this limitation, a high number of independent components can be selected to ensure the detection of the low dose compound. However, it will significantly decrease the quality of the other

calculated signals, associated with the main compounds of the formulation. Moreover, the detection of the low dose compound was mainly based on a spectral comparison between the calculated signal and the pure spectrum of the targeted product. Without knowing the reference pure spectra, identification of the low dose signal might be difficult, if not impossible. Thus, depending on the objective (detection of the main products, detection of a low dose compound, with or without prior knowledge), ICA must be applied carefully by applying the suitable calculation process.

In chapter IV, similar conclusions were highlighted with MCR-ALS which usually applies a reduction of dimensions before the iterative process, with the risk of losing the useful information linked to a low dose compound. By modifying the filtering step, MCR-ALS became an interesting chemometric tool for the detection of a low dose compound in a formulation. However, since constraints have huge impacts on the resolution, there have to be judiciously optimised. The absence/presence maps appeared as very powerful constraint to improve the results. But, to set up these maps, applying methods based on singular value decomposition appeared as inappropriate in the case of a low dose compound. In chapter V, the proposed approach based on orthogonal projections provided good results. However, identification of a compound still requires the use of a threshold during the correlation map interpretations. An inappropriate threshold will lead to false absence/presence maps and hence, will not provide satisfying resolution and a minimum of expertise can be required.

Because hyperspectral imaging uses both spatial and spectral information, two limitations can be considered. First, considering the spatial aspect for a low dose compound, drawback could be associated with the sampling error. Indeed, since the whole sample is not acquired (limited surface or limited depth), pixels containing the low dose product can be missed. It is thus important to acquire a sufficiently large image to ensure the acquisition of all the sample compounds. Second, considering the spectral aspect for a low dose compound with low spectral contributions, the limit of detection will depend on the product. Indeed, a product with very low concentrations, such as an impurity for instance, will not be detected by this approach. Indeed, if the information is not contained in the spectra, multivariate data analysis will not be able to extract it. The proposed methodology could be suitable to other Raman technology which are known to be more sensitive (surface enhanced Raman spectroscopy for instance).

In a future work, several elements should be tested and challenged. Even if a lot of work can be considered to improve the results, the following perspectives were identified as the most interesting studies in the near future:

- Test other ICA algorithms: In this work, the JADE algorithm was applied for model calculations (Chapter III). In order to challenge the presented results, different ICA algorithms such as mutual information based least dependent component analysis (MILCA) [175; 176], SNICA [115] or Fast-ICA [120] for example, should be tested. Even if they do not require a reduction of dimensions, the determination of independence between signals is differently performed and results might be different.

- Initialize MCR-ALS with ICA signals: ICA appeared as an interesting method to calculate pure signals. It has been used in the literature to initialize non-negative matrix factorization [177] and it might be used to calculate initial estimate of the MCR-ALS process. It might provide an alternative approach to usual pure spectrum identification tools (OPA, SIMPLISMA…) when no pure pixels are present in hyperspectral dataset.

- Challenge the results on other datasets: Results of this thesis were only based on Raman hyperspectral imaging of pharmaceutical drug products. However, results and proposed methodologies should be well-adapted to other chemical imaging techniques such as near infrared or matrix-assisted laser desorption ionization imaging or other environments (detection of a contaminant in food engineering)

- Extend the spectral database: In order to ensure the compound detection in an unknown formulation (chapter VI), the spectral database should be expanded with pure compound images. The more the pure compound images in the spectral library, the more efficient the proposed approach.

- Challenge the limit of detection of the method: Limit of detection of the proposed methods should be assessed. Indeed, in this work, we mainly used magnesium stearate, a well-known excipient often used as a lubricant, as the low dose compound because it is present in most pharmaceutical formulations manufactured by direct compression process and because it provides a well-resolved Raman spectrum with sharp Raman peaks. However, the presented work should be tested on other low dose compounds linked to weak modifications of a crystalline form or impurities.

General conclusion

Raman microscopy can be considered as a powerful analytical tool in the pharmaceutical environment to study the distribution of actives and excipients through the entire drug product life cycle. Because the distribution of compounds can modify significantly the quality of the final drug product, the interest of such a technology is growing fast. However, due to the huge amount of data, a direct interpretation of the acquired image is not possible and multivariate data analysis must be applied to extract the targeted information. Using real case examples of pharmaceutical drug products, the objective of the thesis was divided in two main items: i/ Study the compound distributions in a pharmaceutical drug product and ii/ Identify a low dose compound in a sample.

A lot of chemometric tools have been previously applied on chemical imaging dataset to display the distribution of actives and excipients. Most of them are based on variance decomposition or, in a large point of view, on the decomposition of statistical moments. Therefore, some limitations can be observed for the case of a low dose compound, which provides low spatial and spectral contributions in a pharmaceutical drug product.

In the first part of the thesis, this work highlights the potential of independent component analysis and multivariate curve resolution to analyse hyperspectral dataset and extract information of pure compounds. In the case of a low dose compound, the reduction of dimensions or the filtering steps led to a loss of information linked to the targeted product. However, properly used, these two methods appeared as interesting to detect a product with low spatial and spectral contributions. Both algorithms require a high number of components to extract the low dose compound information, which can be mixed with the other compounds of the mixture or spread into noise contributions.

In the second part of the thesis, this work focuses on the signal space, describing the P-dimensional space (one axis per variable) in which the observations can be represented as vectors. By using only the spectral information and orthogonal projections, absence/presence maps of a compound are displayed and are used in the multivariate curve resolution-alternating least squares iterative process. It ensures the detection of a compound without requiring important variations between samples (or pixels). Since it does not require the use of the decomposition of statistical moments on samples, it appears as particularly suitable for the

studied case of a low dose compound. In addition, an iterative approach is proposed to detect pure compounds in a pharmaceutical drug product. Based on a spectral library, spectral distances and orthogonal projections, this approach focuses on the signal space and is also suitable to the detection of a low dose compound.

In the current thesis, the results and proposed methodologies are obtained from Raman microscopy and pharmaceutical drug product. However, it could also be suitable to other hyperspectral dataset including a scarce constituent.

References

[1] Haleem, R. M., Salem, M. Y., Fatahallah, F. A., and Abdelfattah, L. E., "Quality in the pharmaceutical industry - A literature review", *Saudi Pharmaceutical Journal*, 2014.

[2] Roy, J., "7 - The stability of medicines", *An Introduction to Pharmaceutical Sciences Woodhead Publishing Series in Biomedicine*, edited by J. Roy Woodhead Publishing, 2011, pp. 153-181.

[3] Roy, J., "8 - Quality assurance in medicines", *An Introduction to Pharmaceutical Sciences Woodhead Publishing Series in Biomedicine*, edited by J. Roy Woodhead Publishing, 2011, pp. 183-204.

[4] Zhang, C. and Su, J., "Application of near infrared spectroscopy to the analysis and fast quality assessment of traditional Chinese medicinal products", *Acta Pharmaceutica Sinica B*, Vol. 4, No. 3, 2014, pp. 182-192.

[5] De Bleye, C., Chavez, P. F., Mantanus, J., Marini, R., Hubert, P., Rozet, E., and Ziemons, E., "Critical review of near-infrared spectroscopic methods validations in pharmaceutical applications", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 69, 2012, pp. 125-132.

[6] http://www.edqm.eu/

[7] http://www.ema.europa.eu/

[8] Chavez, P. F., Lebrun, P., Sacré, P. Y., De Bleye, C., Netchacovitch, L., Cuypers, S., Mantanus, J., Motte, H., Schubert, M., Evrard, B., Hubert, P., and Ziemons, E., "Optimization of a pharmaceutical tablet formulation based on a design space approach and using vibrational spectroscopy as PAT tool", *International Journal of Pharmaceutics*, Vol. 486, No. 1-2, 2015, pp. 13-20.

[9] Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., and Jent, N., "A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 44, No. 3, 2007, pp. 683-700.

[10] Bloomfield, M., Andrews, D., Loeffen, P., Tombling, C., York, T., and Matousek, P., "Non-invasive identification of incoming raw pharmaceutical materials using Spatially Offset Raman Spectroscopy", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 76, 2013, pp. 65-69.

[11] Lu, F., Weng, X., Chai, Y., Yang, Y., Yu, Y., and Duan, G., "A novel identification system for counterfeit drugs based on portable Raman spectroscopy", *Chemometrics and Intelligent Laboratory Systems*, Vol. 127, 2013, pp. 63-69.

[12] Netchacovitch, L., Thiry, J., De Bleye, C., Chavez, P. F., Krier, F., Sacré, P. Y., Evrard, B., Hubert, P., and Ziemons, E., "Vibrational spectroscopy and microspectroscopy analyzing qualitatively and quantitatively pharmaceutical hot melt extrudates", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 113, 2015, pp. 21-33.

[13] Sabin, G. P., Breitkreitz, M. C., de Souza, A. M., da Fonseca, P., Calefe, L., Moffa, M., and Poppi, R. J., "Analysis of pharmaceutical pellets: An approach using near-infrared chemical imaging", *Analytica Chimica Acta*, Vol. 706, No. 1, 2011, pp. 113-119.

[14] Šašic, S., "Chemical imaging of pharmaceutical granules by Raman global illumination and near-infrared mapping platforms", *Analytica Chimica Acta*, Vol. 611, No. 1, 2008, pp. 73-79.

[15] Gendrin, C., Roggo, Y., and Collet, C., "Pharmaceutical applications of vibrational chemical imaging and chemometrics: A review", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 48, 2008, pp. 533-553.

[16] Ferraro, J. R., Nakamoto, K., and Brown, C. W., "Chapter 1 - Basic Theory", *Introductory Raman Spectroscopy (Second Edition)*, edited by J. R. F. Brown Academic Press, San Diego, 2003, pp. 1-94.

[17] Long, D. A., "Classical Theory of Rayleigh and Raman Scattering", *The Raman Effect*, John Wiley & Sons, Ltd, 2002, pp. 31-48.

[18] Smith, E. and Dent, G., *Modern Raman spectroscopy: a practical approach*, Wiley ed., New York, 2005.

[19] Michelet, A., Boiret, M., Lemhachheche, F., Malec, L., Tfayli, A., and Ziemons E., "Use of Raman spectrometry in the pharmaceutical field", *STP Pharma Pratiques*, Vol. 23, No. 2, 2013, pp. 1-20.

[20] Das, R. S. and Agrawal, Y. K., "Raman spectroscopy: Recent advancements, techniques and applications", *Vibrational Spectroscopy*, Vol. 57, No. 2, 2011, pp. 163-176.

[21] McCreery, R. L., *Raman Spectroscopy for Chemical Analysis*, New York, 2005.

[22] Ozaki, Y. and Šašic, S., "Introduction to Raman Spectroscopy", *Pharmaceutical Applications of Raman Spectroscopy*, John Wiley & Sons, Inc., 2007, pp. 1-28.

[23] Lyndgaard, L. B., van den Berg, F., and de Juan, A., "Quantification of paracetamol through tablet blister packages by Raman spectroscopy and multivariate curve resolution-alternating least squares", *Chemometrics and Intelligent Laboratory Systems*, Vol. 125, 2013, pp. 58-66.

[24] Pelletier, M., *Analytical Applications of Raman Spectroscopy*, Wiley-Blackwell ed., Oxford, 1999.

[25] Bakeev, K. A., *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries*, *2nd Edition*, Wiley ed., Chichester, 2010.

[26] Griffiths, P. R. and Miseo, E. V., "Infrared and Raman Instrumentation for Mapping and Imaging", *Infrared and Raman Spectroscopic Imaging*, Wiley-VCH Verlag GmbH & Co. KGaA, 2014, pp. 1-56.

[27] Amigo, J. M., Babamoradi, H., and Elcoroaristizabal, S., "Hyperspectral image analysis. A tutorial", *Analytica Chimica Acta*, 2015.

[28] Šašic, S., *Pharmaceutical Applications of Raman Spectroscopy*, Wiley ed. 2007.

[29] Wartewig, S. and Neubert, R. H. H., "Pharmaceutical applications of Mid-IR and Raman spectroscopy", *Advanced Drug Delivery Reviews*, Vol. 57, No. 8, 2005, pp. 1144-1170.

[30] Gowen, A. A., O'Donnell, C. P., Cullen, P. J., and Bell, S. E. J., "Recent applications of Chemical Imaging to pharmaceutical process monitoring and quality control", *European Journal of Pharmaceutics and Biopharmaceutics*, Vol. 69, 2008, pp. 10-22.

[31] Vankeirsbilck, T., Vercauteren, A., Baeyens, W., Van der Weken, G., Verpoort, F., Vergote, G., and Remon, J. P., "Applications of Raman spectroscopy in pharmaceutical analysis", *TrAC Trends in Analytical Chemistry*, Vol. 21, No. 12, 2002, pp. 869-877.

[32] Griffen, J., Owen, A., and Matousek, P., "Comprehensive Quantification of Tablets with Multiple Active Pharmaceutical Ingredients using Transmission Raman Spectroscopy - A proof of concept study", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 115, 2015, pp. 277-282.

[33] Zhang, Y. and McGeorge, G., "Quantitative Analysis of Pharmaceutical Bilayer Tablets Using Transmission Raman Spectroscopy", *Journal of Pharmaceutical Innovation*, Vol. 10, No. 3, 2015, pp. 269-280.

[34] Skorda, D. and Kontoyannis, C. G., "Identification and quantitative determination of atorvastatin calcium polymorph in tablets using FT-Raman spectroscopy", *Talanta*, Vol. 74, No. 4, 2008, pp. 1066-1070.

[35] Hu, Y., Wikström, H., Byrn, S. R., and Taylor, L. S., "Estimation of the transition temperature for an enantiotropic polymorphic system from the transformation kinetics monitored using Raman spectroscopy", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 45, No. 4, 2007, pp. 546-551.

[36] Roggo, Y., Degardin, K., and Margot, P., "Identification of pharmaceutical tablets by Raman spectroscopy and chemometrics", *Talanta*, Vol. 81, No. 3, 2010, pp. 988-995.

[37] Dégardin, K., Roggo, Y., Been, F., and Margot, P., "Detection and chemical profiling of medicine counterfeits by Raman spectroscopy and chemometrics", *Analytica Chimica Acta*, Vol. 705, No. 1-2, 2011, pp. 334-341.

[38] Eliasson, C. and Matousek, P., "Noninvasive Authentication of Pharmaceutical Products through Packaging Using Spatially Offset Raman Spectroscopy", *Analytical Chemistry*, Vol. 79, No. 4, 2007, pp. 1696-1701.

[39] De Beer, T. R. M., Bodson, C., Dejaegher, B., Walczak, B., Vercruysse, P., Burggraeve, A., Lemos, A., Delattre, L., Heyden, Y. V., Remon, J. P., Vervaet, C., and Baeyens, W. R. G., "Raman spectroscopy as a process analytical technology (PAT) tool for the in-line monitoring and understanding of a powder blending process", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 48, No. 3, 2008, pp. 772-779.

[40] De Beer, T. R. M., Baeyens, W. R. G., Ouyang, J., Vervaet, C., and Remon, J. P., "Raman spectroscopy as a process analytical technology tool for the understanding and the quantitative in-line monitoring of the homogenization process of a pharmaceutical suspension", *Analyst*, Vol. 131, No. 10, 2006, pp. 1137-1144.

[41] Islam, M. T., Rodriguez-Hornedo, N., Ciotti, S., and Ackermann, C., "The Potential of Raman Spectroscopy as a Process Analytical Technique During Formulations of Topical Gels and Emulsions", *Pharmaceutical Research*, Vol. 21, No. 10, 2004, pp. 1844-1851.

[42] Kauffman, J. F., Dellibovi, M., and Cunningham, C. R., "Raman spectroscopy of coated pharmaceutical tablets and physical models for multivariate calibration to tablet coating thickness", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 43, No. 1, 2007, pp. 39-48.

[43] Palou, A., Cruz, J., Blanco, M., Tomás, J., de los Rios, J. n., and Alcalá, M., "Determination of drug, excipients and coating distribution in pharmaceutical tablets using NIR-CI", *Journal of Pharmaceutical Analysis*, Vol. 2, No. 2, 2012, pp. 90-97.

[44] Tres, F., Treacher, K., Booth, J., Hughes, L. P., Wren, S. A. C., Aylott, J. W., and Burley, J. C., "Real time Raman imaging to understand dissolution performance of amorphous solid dispersions", *Journal of Controlled Release*, Vol. 188, 2014, pp. 53-60.

[45] Paudel, A., Raijada, D., and Rantanen, J., "Raman spectroscopy in pharmaceutical product design", *Advanced Drug Delivery Reviews*, Vol. 89, 2015, pp. 3-20.

[46] de Souza Lins Borba, F., Saldanha Honorato, R., and de Juan, A., "Use of Raman spectroscopy and chemometrics to distinguish blue ballpoint pen inks", *Forensic Science International*, Vol. 249, 2015, pp. 73-82.

[47] Reddy, R. K. and Bhargava, R., "Chemometric Methods for Biomedical Raman Spectroscopy and Imaging", *Emerging Raman Applications and Techniques in Biomedical and Pharmaceutical Fields*, edited by P. Matousek and M. D. Morris Biological and Medical Physics, Biomedical Engineering, Springer Berlin Heidelberg, 2010, pp. 179-213.

[48] Zhang, L., Henson, M. J., and Sekulic, S. S., "Multivariate data analysis for Raman imaging of a model pharmaceutical tablet", *Analytica Chimica Acta*, Vol. 545, No. 2, 2005, pp. 262-278.

[49] Amigo, J. M., Cruz, J., Bautista, M., Maspoch, S., Coello, J., and Blanco, M., "Study of pharmaceutical samples by NIR chemical-image and multivariate analysis", *TrAC Trends in Analytical Chemistry*, Vol. 27, No. 8, 2008, pp. 696-713.

[50] Lopes, M. B., Wolff, J. C., Bioucas Dias, J. M., and Figueiredo, M. A. T., "Determination of the composition of counterfeit Heptodin tablets by near infrared chemical imaging and classical least squares estimation", *Analytica Chimica Acta*, Vol. 641, No. 1–2, 2009, pp. 46-51.

[51] Puchert, T., Lochmann, D., Menezes, J. C., and Reich, G., "Near-infrared chemical imaging (NIR-CI) for counterfeit drug identificationΓÇöA four-stage concept with a novel approach of data processing (Linear Image Signature)", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 51, No. 1, 2010, pp. 138-145.

[52] Vajna, B., Farkas, A., Pataki, H., Zsigmond, Z., Igricz, T., and Marosi, G., "Testing the performance of pure spectrum resolution from Raman hyperspectral images of differently manufactured pharmaceutical tablets", *Analytica Chimica Acta*, Vol. 712, 2012, pp. 45-55.

[53] Amigo, J. M., Ravn, C., Gallagher, N. B., and Bro, R., "A comparison of a common approach to partial least squares-discriminant analysis and classical least squares in hyperspectral imaging", *International Journal of Pharmaceutics*, Vol. 373, No. 1–2, 2009, pp. 179-182.

[54] Ozeki, Y., Umemura, W., Otsuka, Y., Satoh, S., Hashimoto, H., Sumimura, K., Nishizawa, N., Fukui, K., and Itoh, K., "High-speed molecular spectral imaging of tissue with stimulated Raman scattering", *Nature Photonics*, Vol. 6, No. 12, 2012, pp. 845-851.

[55] Gonzales, R. C. and Woods, E., *Digital image processing*, 3rd edition ed., Prentice Hall 2007.

[56] Rinnan, A., "Pre-processing in vibrational spectroscopy - when, why and how", *Analytical Methods*, Vol. 6, No. 18, 2014, pp. 7124-7129.

[57] Zeaiter, M., Roger, J. M., and Bellon-Maurel, V., "Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods", *TrAC Trends in Analytical Chemistry*, Vol. 24, No. 5, 2005, pp. 437-445.

[58] Reisner, L. A., Cao, A., and Pandya, A. K., "An integrated software system for processing, analyzing, and classifying Raman spectra", *Chemometrics and Intelligent Laboratory Systems*, Vol. 105, No. 1, 2011, pp. 83-90.

[59] Savitzky, A. and Golay, M. J. E., "Smoothing and Differentiation of Data by Simplified Least Squares Procedures", *Analytical Chemistry*, Vol. 36, 1964, pp. 1627-1639.

[60] Mozharov, S., Nordon, A., Littlejohn, D., and Marquardt, B., "Automated Cosmic Spike Filter Optimized for Process Raman Spectroscopy", *Applied Spectroscopy*, Vol. 66, No. 11, 2012, pp. 1326-1333.

[61] Post Sabin, G., de Souza, A. M., Breitkreitz, M. C., and Poppi, R. J., "Development of an algorithm for identification and correction of spikes in raman imaging spectroscopy", *Quimica Nova*, Vol. 35, 2012, pp. 612-615.

[62] Zhang, L. and Henson, M. J., "A practical algorithm to remove cosmic spikes in Raman imaging data for pharmaceutical applications", *Applied Spectroscopy*, Vol. 61, No. 9, 2007, pp. 1015-1020.

[63] Behrend, C. J., Tarnowski, C. P., and Morris, M. D., "Identification of Outliers in Hyperspectral Raman Image Data by Nearest Neighbor Comparison", *Applied Spectroscopy*, Vol. 56, No. 11, 2002, pp. 1458-1461.

[64] Jirasek, A., Schulze, G., Yu, M. M. L., Blades, M. W., and Turner, R. F. B., "Accuracy and Precision of Manual Baseline Determination", *Applied Spectroscopy*, Vol. 58, No. 12, 2004, pp. 1488-1499.

[65] Zhao, J., Lui, H., McLean, D. I., and Zeng, H., "Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy", *Applied Spectroscopy*, Vol. 61, No. 11, 2007, pp. 1225-1232.

[66] Lieber, C. A. and Mahadevan-Jansen, A., "Automated Method for Subtraction of Fluorescence from Biological Raman Spectra", *Applied Spectroscopy*, Vol. 57, No. 11, 2003, pp. 1363-1367.

[67] Prakash, B. D. and Wei, Y. C., "A fully automated iterative moving averaging (AIMA) technique for baseline correction", *Analyst*, Vol. 136, No. 15, 2011, pp. 3130-3135.

[68] Eilers, P. H. C., "Parametric Time Warping", *Analytical Chemistry*, Vol. 76, No. 2, 2004, pp. 404-411.

[69] Eilers, P. H. C., "A Perfect Smoother", *Analytical Chemistry*, Vol. 75, No. 14, 2003, pp. 3631-3636.

[70] Peng, J., Peng, S., Jiang, A., Wei, J., Li, C., and Tan, J., "Asymmetric least squares for multiple spectra baseline correction", *Analytica Chimica Acta*, Vol. 683, No. 1, 2010, pp. 63-68.

[71] Rinnan, A., Berg, F. v. d., and Engelsen, S. B., "Review of the most common pre-processing techniques for near-infrared spectra", *TrAC Trends in Analytical Chemistry*, Vol. 28, No. 10, 2009, pp. 1201-1222.

[72] Barnes, R. J., Dhanoa, M. S., and Lister, S. J., "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra", *Applied Spectroscopy*, Vol. 43, 1989, pp. 772-777.

[73] Martens, H. and Stark, E., "Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 9, No. 8, 1991, pp. 625-635.

[74] Shaver, J., "Chemometrics for Raman spectroscopy", *Handbook of Raman Spectroscopy: From the Research Laboratory to the Process Line*, edited by I. R. Lewis and H. G. M. Edwards New York, 2001, pp. 275-306.

[75] Wold, S., Esbensen, K., and Geladi, P., "Proceedings of the Multivariate Statistical Workshop for Geologists and GeochemistsPrincipal component analysis", *Chemometrics and Intelligent Laboratory Systems*, Vol. 2, No. 1, 1987, pp. 37-52.

[76] Wu, W., Massart, D. L., and de Jong, S., "The kernel PCA algorithms for wide data. Part I: Theory and algorithms", *Chemometrics and Intelligent Laboratory Systems*, Vol. 36, No. 2, 1997, pp. 165-172.

[77] Gowen, A. A., O'Donnell, C. P., Taghizadeh, M., Cullen, P. J., Frias, J. M., and Downey, G., "Hyperspectral imaging combined with principal component analysis for bruise damage detection on white mushrooms (Agaricus bisporus)", *Journal of Chemometrics*, Vol. 22, No. 3-4, 2008, pp. 259-267.

[78] Berthiaux, H., Mosorov, V., Tomczak, L., Gatumel, C., and Demeyre, J. F., "Principal component analysis for characterising homogeneity in powder mixing using image processing techniques", *Chemical Engineering and Processing: Process Intensification*, Vol. 45, No. 5, 2006, pp. 397-403.

[79] Stone, J. V., *Independent component analysis – A tutorial introduction*, A Bradford Book ed., London, 2004.

[80] De Lathauwer, L., De Moor, B., and Vandewalle, J., "An introduction to independent component analysis", *Journal of Chemometrics*, Vol. 14, No. 3, 2000, pp. 123-149.

[81] de Juan, A. and Tauler, R., "Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution", *Analytica Chimica Acta*, Vol. 500, 2003, pp. 195-210.

[82] Jaumot, J., de Juan, A., and Tauler, R., "MCR-ALS GUI 2.0: New features and applications", *Chemometrics and Intelligent Laboratory Systems*, Vol. 140, 2015, pp. 1-12.

[83] de Juan, A., Tauler, R., Dyson, R., Marcolli, C., Rault, M., and Maeder, M., "Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis", *TrAC Trends in Analytical Chemistry*, Vol. 23, No. 1, 2004, pp. 70-79.

[84] de Juan, A., Piqueras, S., Maeder, M., Hancewicz, T., Duponchel, L., and Tauler, R., "Chemometric Tools for Image Analysis", *Infrared and Raman Spectroscopic Imaging*, Wiley-VCH Verlag GmbH & Co. KGaA, 2014, pp. 57-110.

[85] Hugelier, S., Devos, O., and Ruckebusch, C., "On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis", *Journal of Chemometrics*, Vol. 29, No. 10, 2015, pp. 557-561.

[86] Zhang, X., Juan, A. d., and Tauler, R., "Local rank-based spatial information for improvement of remote sensing hyperspectral imaging resolution", *Talanta*, Vol. 146, 2016, pp. 1-9.

[87] Windig, W. and Guilment, J., "Interactive self-modeling mixture analysis", *Analytical Chemistry*, Vol. 63, No. 14, 1991, pp. 1425-1432.

[88] Gourvénec, S., Lamotte, C., Pestiaux, P., and Massart, D. L., "Use of the Orthogonal Projection Approach (OPA) to Monitor Batch Processes", *Applied Spectroscopy*, Vol. 57, No. 1, 2003, pp. 80-87.

[89] Boiret, M., Rutledge, D. N., Gorretta, N., Ginot, Y. M., and Roger, J. M., "Application of independent component analysis on Raman images of a pharmaceutical drug product: Pure spectra determination and spatial distribution of constituents", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 90, 2014, pp. 78-84.

[90] Keller, H. R. and Massart, D. L., "Evolving factor analysis", *Chemometrics and Intelligent Laboratory Systems*, Vol. 12, No. 3, 1991, pp. 209-224.

[91] Bell, S. E. J., Beattie, J. R., McGarvey, J. J., Peters, K. L., Sirimuthu, N. M. S., and Speers, S. J., "Development of sampling methods for Raman analysis of solid dosage forms of therapeutic and illicit drugs", *Journal of Raman Spectroscopy*, Vol. 35, 2004, pp. 409-417.

[92] Šašic, S. and Whitlock, M., "Raman Mapping of Low-Content Active-Ingredient Pharmaceutical Formulations. Part II: Statistically Optimized Sampling for Detection of Less Than 1% of an Active Pharmaceutical Ingredient", *Applied Spectroscopy*, Vol. 62, No. 8, 2008, pp. 916-921.

[93] Šašic, S. and Mehrens, S., "Raman chemical mapping of low-content active pharmaceutical ingredient formulations. III. Statistically optimized sampling and detection of polymorphic forms in tablets on stability", *Analytical Chemistry*, Vol. 84, 2012, pp. 1019-1025.

[94] Hausman, D. S., Cambron, R. T., and Sakr, A., "Application of Raman spectroscopy for on-line monitoring of low dose blend uniformity", *International Journal of Pharmaceutics*, Vol. 298, No. 1, 2005, pp. 80-90.

[95] Šašic, S., "Raman Mapping of Low-Content API Pharmaceutical Formulations. I. Mapping of Alprazolam in Alprazolam/Xanax Tablets", *Pharmaceutical Research*, Vol. 24, No. 1, 2007, pp. 58-65.

[96] Ziemons, E., Mantanus, J., Lebrun, P., Rozet, E., Evrard, B., and Hubert, P., "Acetaminophen determination in low-dose pharmaceutical syrup by NIR

spectroscopy", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 53, No. 3, 2010, pp. 510-516.

[97] Chalus, P., Roggo, Y., Walter, S., and Ulmschneider, M., "Near-infrared determination of active substance content in intact low-dosage tablets", *Talanta*, Vol. 66, No. 5, 2005, pp. 1294-1302.

[98] Li, B., Calvet, A., Casamayou-Boucau, Y., Morris, C., and Ryder, A. G., "Low-Content Quantification in Powders Using Raman Spectroscopy: A Facile Chemometric Approach to Sub 0.1% Limits of Detection", *Analytical Chemistry*, Vol. 87, No. 6, 2015, pp. 3419-3428.

[99] Hennigan, M. C. and Ryder, A. G., "Quantitative polymorph contaminant analysis in tablets using Raman and near infra-red spectroscopies", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 72, 2013, pp. 163-171.

[100] Porfire, A., Rus, L., Vonica, A. L., and Tomuta, I., "High-throughput NIR-chemometric methods for determination of drug content and pharmaceutical properties of indapamide powder blends for tabletting", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 70, 2012, pp. 301-309.

[101] Alcalá, M., Léon, J., Ropero, J., Blanco, M., and Romanach, R. J., "Analysis of low content drug tablets by transmission near infrared spectroscopy: Selection of calibration ranges according to multivariate detection and quantitation limits of PLS models", *Journal of Pharmaceutical Sciences*, Vol. 97, No. 12, 2008, pp. 5318-5327.

[102] Ferré, J. and Faber, N., "Net analyte signal calculation for multivariate calibration", *Chemometrics and Intelligent Laboratory Systems*, Vol. 69, 2003, pp. 123-136.

[103] Lorber, A., "Net analyte signal calculation in multivariate calibration", *Analytical Chemistry*, Vol. 69, 1997, pp. 1620-1626.

[104] Blanco, M., Castillo, M., Peinado, A., and Beneyto, R., "Determination of low analyte concentrations by near-infrared spectroscopy: Effect of spectral pretreatments and estimation of multivariate detection limits", *Analytica Chimica Acta*, Vol. 581, No. 2, 2007, pp. 318-323.

[105] Rantanen, J., "Process analytical applications of Raman spectroscopy", *Journal of pharmacy and pharmacology*, Vol. 59, 2007, pp. 171-177.

[106] Widjaja, E., Kanaujia, P., Lau, G., Ng, W. K., Garland, M., Saal, C., Hanefeld, A., Fischbach, M., Maio, M., and Tan, R. B. H., "Detection of trace cristallinity in an amorphous system using Raman microscopy and chemometrics analysis", *European Journal of Pharmaceutical Sciences*, Vol. 42, 2011, pp. 45-54.

[107] Matousek, P. and Parker, A. W., "Non-invasive probing of pharmaceutical capsules using transmission Raman spectroscopy", *Journal of Raman Spectroscopy*, Vol. 38, 2007, pp. 563-567.

[108] Zhang, X. and Tauler, R., "Application of Multivariate Curve Resolution Alternative Least Squares (MCR-ALS) to remote sensing hyperspectral imaging", *Analytica Chimica Acta*, Vol. 762, 2013, pp. 25-38.

[109] Xiabo, Z., Jiewen, Z., Holmes, M., Hanpin, M., Jiyong, S., Xiaopin, Y., and Yanxiao, L., "Independent component analysis in information extraction from visible/near-infrared

hyperspectral imaging data of cucumber leaves", *Chemometrics and Intelligent Laboratory Systems*, Vol. 104, 2010, pp. 265-270.

[110] Roggo, Y., Edmond, A., Chalus, P., and Ulmschneider, M., "Infrared hyperspectral imaging for qualitative analysis of pharmaceutical solid forms", *Analytica Chimica Acta*, Vol. 535, 2005, pp. 79-87.

[111] Goetz, M. J., Coté, G. L., Erckens, R., March, W., and Motamedi, M., "Application of a multivariate technique to Raman spectra for quantification of body chemicals", *IEEE Transactions on Biomedical Engineering*, Vol. 42, 1995, pp. 728-731.

[112] Grahn, H. and Geladi, P., *Techniques and Applications of Hyperspectral Image Analysis*, John Wiley & son Ltd ed., New York, 2007.

[113] Lopes, M. B. and Wolff, J. C., "Investigation into classification/sourcing of suspect counterfeit Heptodin™ tablets by near infrared chemical imaging", *Analytica Chimica Acta*, Vol. 633, 2009, pp. 149-155.

[114] Vajna, B., Farkas A., Pataki, H., Zsigmond, Z., Igricz, T., and Marosi, G., "Testing the performance of pure spectrum resolution from Raman hyperspectral images of differently manufactured pharmaceutical tablets", *Analytica Chimica Acta*, Vol. 712, 2012, pp. 45-55.

[115] Monakhova, Y. B., stakhov, S. A., Kraskov, A., and Mushtakova, S. P., "Independent components in spectroscopic analysis of complex mixtures", *Chemometrics and Intelligent Laboratory Systems*, Vol. 103, 2010, pp. 108-115.

[116] Wang, G., Ding, Q., and Hou, Z., "Independent component analysis and its applications in signal processing for analytical chemistry", *Trends in Analytical Chemistry*, Vol. 27, 2008, pp. 368-376.

[117] Hyvärinen, A. and Oja, E., "Independent component analysis: algorithms and applications", *Neural Networks*, Vol. 13, 2000, pp. 411-430.

[118] Lin, H., Marjanovic, O., Lennox, B., Šašic, S., and Clegg, I. M., "Multivariate statistical analysis of Raman images of a pharmaceutical tablet", *Applied Spectroscopy*, Vol. 66, 2012, pp. 272-281.

[119] de Lathauwer, L., de Moor, B., and Vandewalle, J., "An introduction to independent component analysis", *Journal of Chemometrics*, Vol. 14, 2000, pp. 123-149.

[120] Hyvärinen, A. and Oja, E., "A fast fixed-point algorithm for independent component analysis", *Neural Computation*, Vol. 9, 1997, pp. 1483-1492.

[121] Learned-Miller, E. G. and Fisher, J. W., "ICA using spacings estimates of entropy", *Journal of Machine Learning Research*, Vol. 4, 2003, pp. 1271-1295.

[122] Cardoso, J. F., "High-order contrasts for independent component analysis", *Neural Computation*, Vol. 11, 1999, pp. 157-192.

[123] Rutledge, D. N. and Jouan-Rimbaud Bouveresse, D., "Independent Component Analysis with the JADE algorithm", *Trends in Analytical Chemistry*, Vol. 50, 2013, pp. 22-32.

[124] Jouan-Rimbaud Bouveresse, D., Moya-Gonzales, A., Ammari, F., and Rutledge, D. N., "Two novel methods for the determination of the number of components in independent

component analysis models", *Chemometrics and Intelligent Laboratory Systems*, Vol. 112, 2012, pp. 24-32.

[125] http://perso.telecom-paristech.fr/~cardoso/Algo/Jade/jadeR.m

[126] Jouan-Rimbaud Bouveresse, D., Benabid, H., and Rutledge, D. N., "independent component analysis as a pretreatment method for parallel factor analysis to eliminate artefacts from multiway data", *Analytica Chimica Acta*, Vol. 589, 2007, pp. 216-224.

[127] Kazarian, S. and Higgins, J., "A closer look at polymers", *Chemistry and Industry*, Vol. 10, 2002, pp. 21-23.

[128] Koljenovic, S., Bakker Schut, T. C., Wolthuis, R., de Jong, B., Santos, L., Caspers, P. J., Kros, J. M., and Puppels, G. J., "Tissue characterization using high wave number Raman spectroscopy.", *Journal of Biomedical Optics*, Vol. 10, No. 3, 2005, pp. 031116.

[129] Bautista, M. and Cruz, J. B. M., "Study of component distribution in pharmaceutical binary powder mixtures by near infrared chemical imaging", *Journal of Spectral Imaging*, Vol. 3, 2012, pp. 1-9.

[130] Schönbichler, S. A., Bittner, L. K. H., Weiss, A. K. H., Griesser, U. J., Pallua, J. D., and Huck, C. W., "Comparison of NIR chemical imaging with conventional NIR, Raman and ATR-IR spectroscopy for quantification of furosemide crystal polymorphs in ternary powder mixtures", *European Journal of Pharmaceutics and Biopharmaceutics*, Vol. 84, No. 3, 2013, pp. 616-625.

[131] Kwok, K., "Analysis of Cialis tablets using Raman microscopy and multivariate curve resolution", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 66, 2012, pp. 126-135.

[132] Vajna, B., Patyi, G., Nagy, Z., Bodis, A., Farkas, A., and Marosi, G., "Comparison of chemometric methods in the analysis of pharmaceuticals with hyperspectral Raman imaging", *Journal of Raman Spectroscopy*, Vol. 42, No. 11, 2011, pp. 1977-1986.

[133] Clarke, F., "Extracting process-related information from pharmaceutical dosage forms using near infrared microscopy", *Vibrational Spectroscopy*, Vol. 34, No. 1, 2004, pp. 25-35.

[134] Šašic, S. and Clark, D. A., "Defining a strategy for chemical imaging of industrial pharmaceutical samples on Raman line-mapping and global illumination instruments", *Applied Spectroscopy*, Vol. 60, 2006, pp. 494-502.

[135] Burger, J. and Geladi, P., "Hyperspectral NIR image regression part II: dataset preprocessing diagnostics", *Journal of Chemometrics*, Vol. 20, No. 3-4, 2006, pp. 106-119.

[136] Furukawa, T., Sato, H. S. H., Noda, I., and Ochiai, S., "Evaluation of homogeneity of binary blends of poly(3-hydroxybutyrate) and poly(L-lactic acid) studied by near infrared chemical imaging (NIRCI)", *Analytical Sciences*, Vol. 23, 2007, pp. 871-876.

[137] Piqueras, S., Duponchel, L., Tauler, R., and de Juan, A., "Resolution and segmentation of hyperspectral biomedical images by Multivariate Curve Resolution-Alternating Least Squares", *Analytica Chimica Acta*, Vol. 705, 2011, pp. 182-192.

[138] Gendrin, C., Roggo, Y., and Collet, C., "Content uniformity of pharmaceutical solid dosage forms by near infrared hyperspectral imaging: A feasibility study", *Talanta*, Vol. 73, No. 4, 2007, pp. 733-741.

[139] Abdollahi, H. and Tauler, R., "Uniqueness and rotation ambiguities in Multivariate Curve Resolution methods", *Chemometrics and Intelligent Laboratory Systems*, Vol. 108, No. 2, 2011, pp. 100-111.

[140] Jaumot, J., Gargallo, R., de Juan, A., and Tauler, R., "A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB", *Chemometrics and Intelligent Laboratory Systems*, Vol. 76, No. 1, 2005, pp. 101-110.

[141] de Juan, A., Maeder, M., Hancewicz, T., Duponchel, L., and Tauler, R., "Chemometric Tools for Image Analysis", *Infrared and Raman Spectroscopic Imaging*, Wiley-VCH Verlag GmbH & Co. KGaA, 2009, pp. 65-109.

[142] Lee, E., "Raman Spectral Imaging on Pharmaceutical Products", *Infrared and Raman Spectroscopic Imaging*, Wiley-VCH Verlag GmbH & Co. KGaA, 2009, pp. 377-402.

[143] Wang, J., Wen, H., and Desai, D., "Lubrication in tablet formulations", *European Journal of Pharmaceutics and Biopharmaceutics*, Vol. 75, No. 1, 2010, pp. 1-15.

[144] de Juan, A., Maeder, M., Hancewicz, T., and Tauler, R., "Use of local rank-based spatial information for resolution of spectroscopic images", *Journal of Chemometrics*, Vol. 22, 2008, pp. 291-298.

[145] de Juan, A., Maeder, M., Hancewicz, T., and Tauler, R., "Local rank analysis for exploratory spectroscopic image analysis. Fixed Size Image Window-Evolving Factor Analysis", *Chemometrics and Intelligent Laboratory Systems*, Vol. 77, 2005, pp. 64-74.

[146] Geladi, P. and Grahn, H., *Multivariate image analysis in chemistry and related areas: chemometrics image analysis*, John Wiley & Sons ed., Chichester, 1996.

[147] Tauler, R., Maeder, M., and de Juan, A., "2.24 - Multiset Data Analysis: Extended Multivariate Curve Resolution", *Comprehensive Chemometrics*, edited by S. D. B. Walczak Elsevier, Oxford, 2009, pp. 473-505.

[148] Piqueras, S., Burger, J., Tauler, R., and de Juan, A., "Relevant aspects of quantification and sample heterogeneity in hyperspectral image resolution", *Chemometrics and Intelligent Laboratory Systems*, Vol. 117, 2012, pp. 169-182.

[149] Boiret, M., de Juan, A., Gorretta, N., Ginot, Y. M., and Roger, J. M., "Distribution of a low dose compound within pharmaceutical tablet by using multivariate curve resolution on Raman hyperspectral images", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 103, 2015, pp. 35-43.

[150] Sarraguça, M. C. and Lopes, J. A., "The use of net analyte signal (NAS) in near infrared spectroscopy pharmaceutical applications: Interpretability and figures of merit", *Analytica Chimica Acta*, Vol. 642, 2009, pp. 179-185.

[151] Lorber, A., "Error propagation and figures of merit for quantification by solving matrix equations", *Analytical Chemistry*, Vol. 58, 1986, pp. 1167-1172.

[152] Boulet, J. C. and Roger, J. M., "Pretreatments by means of orthogonal projections", *Chemometrics and Intelligent Laboratory Systems*, Vol. 117, 2012, pp. 61-69.

[153] Goicoechea, H. C. and Olivieri, A. C., "A comparison of orthogonal signal correction and net analyte preprocessing methods. Theoretical and experimental study", *Chemometrics and Intelligent Laboratory Systems*, Vol. 56, No. 2, 2001, pp. 73-81.

[154] Jaillais, B., Boulet, J. C., Roger, J. M., Balfourier, F., Berbezy, P., and Bertrand, D., "Application of direct calibration in multivariate image analysis of heterogeneous materials", *Analytica Chimica Acta*, Vol. 734, 2012, pp. 45-53.

[155] Fini, G., "Applications of Raman spectroscopy to pharmacy", *Journal of Raman Spectroscopy*, Vol. 35, No. 5, 2004, pp. 335-337.

[156] Amigo, J. M., "Practical issues of hyperspectral imaging analysis of solid dosage forms", *Analytical and Bioanalytical Chemistry*, Vol. 398, No. 1, 2010, pp. 93-109.

[157] Smith, G. P. S., McGoverin, C. M., Fraser, S. J., and Gordon, K. C., "Raman imaging of drug delivery systems", *Advanced Drug Delivery Reviews*, Vol. 89, 2015, pp. 21-41.

[158] Fortunato de Carvalho Rocha, W., Sabin, G. P., Março, P. H., and Poppi, R. J., "Quantitative analysis of piroxicam polymorphs pharmaceutical mixtures by hyperspectral imaging and chemometrics", *Chemometrics and Intelligent Laboratory Systems*, Vol. 106, No. 2, 2011, pp. 198-204.

[159] Wu, J., "Linear quantification calibration of crystallinity at subpercent and its evaluation based on spectral and spatial information inherited in Raman chemical images", *Journal of Raman Spectroscopy*, Vol. 45, 2014, pp. 686-695.

[160] Doub, W. H., Adams, W. P., Spencer, J. A., Buhse, L. F., Nelson, M. P., and Treado, P. J., "Raman Chemical Imaging for Ingredient-specific Particle Size Characterization of Aqueous Suspension Nasal Spray Formulations: A Progress Report", *Pharmaceutical Research*, Vol. 24, No. 5, 2007, pp. 934-945.

[161] Sacré, P. Y., Lebrun, P., Chavez, P. F., Bleye, C. D., Netchacovitch, L., Rozet, E., Klinkenberg, R., Streel, B., Hubert, P., and Ziemons, E., "A new criterion to assess distributional homogeneity in hyperspectral images of solid pharmaceutical dosage forms", *Analytica Chimica Acta*, Vol. 818, 2014, pp. 7-14.

[162] Martinez, P. J., Pérez, R. M., Plaza, A., Aguilar, P. L., Cantero, M. C., and Plaza, J., "Endmember extraction algorithms from hyperspectral images", *Annals of Geophysics*, Vol. 49, No. 1, 2006, pp. 93-101.

[163] Veganzones, M. A. and Grana, M., "Endmember Extraction Methods: A Short Review", *Knowledge-Based Intelligent Information and Engineering Systems*, edited by I. Lovrek, R. Howlett, and L. Jain Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2008, pp. 400-407.

[164] Boardman, J. W., Kruse, F. A., and Greem, R. Q., "Mapping target signature via partial unmixing of AVIRIS data", in *Summaries of the 5th JPL Airborne Earth Science Workshop*, 1995.

[165] Plaza, A., Martinez, P., Pérez, R. M., and Plaza, J., "Spatial/spectral endmember extraction by multi-dimensional morphological operations", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 40, No. 9, 2002, pp. 2025-2041.

[166] Winter, M. E., "N-FINDR: an algorithm for fast autonomous spectral endmember determination in hyperspectral data", *SPIE Proceedings*, Vol. 3753, 1999, pp. 266-275.

[167] Nascimento, J. and Bioucas Dias, J. M., "Vertex component analysis: a fast algorithm to unmix hyperspectral data", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 43, 2005, pp. 898-910.

[168] Widjaja, E. and Seah, R. K. H., "Application of Raman microscopy and band-target entropy minimization to identify minor components in model pharmaceutical tablets", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 46, No. 2, 2008, pp. 274-281.

[169] Sacré, P. Y., Deconinck, E., Saerens, L., de Beer, T., Courselle, P., Vancauwenberghe, R., Chiap, P., Crommen, J., and De Beer, J. O., "Detection of counterfeit Viagra by Raman microspectroscopy imaging and multivariate analysis", *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 56, No. 2, 2011, pp. 454-461.

[170] Vajna, B., Pataki, H., Nagy, Z., Farkas, I., and Marosi, G., "Characterization of melt extruded and conventional Isoptin formulations using Raman chemical imaging and chemometrics", *International Journal of Pharmaceutics*, Vol. 419, 2011, pp. 107-113.

[171] de Veij, M., Vandenabeele, P., de Beer, T., Remon, J.-P., and Moens, L., "Reference database of Raman spectra of pharmaceutical excipients", *Journal of Raman Spectroscopy*, Vol. 40, 2008, pp. 297-307.

[172] Mathieu Boiret , Anna de Juan, Nathalie Gorretta, Yves-Michel Ginot, and Jean-Michel Roger, "Local rank constraints by orthogonal projections for image resolution analysis: application to the determination of a low dose pharmaceutical compound", *Analytica Chimica Acta*, 2015.

[173] Kruse, F. A., Lefkoff, A. B., Boardman, J. W., Heidebrecht, K. B., Shapiro, A. T., Barloon, P. J., and Goetz, A. F. H., "The spectral image processing system (SIPS) – interactive visualization and analysis of imaging spectrometer data", *Remote Sensing of Environment*, Vol. 44, No. 2-3, 1993, pp. 145-163.

[174] Tukey, J. W., *Exploratory data analysis*, First edition ed., Addison-Wesley, Reading, MA, 1977.

[175] Monakhova, Y. B., Tsikin, A. M., Mushtakova, S. P., and Mecozzi, M., "Independent component analysis and multivariate curve resolution to improve spectral interpretation of complex spectroscopic data sets: Application to infrared spectra of marine organic matter aggregates", *Microchemical Journal*, Vol. 118, 2015, pp. 211-222.

[176] Stögbauer, H., Kraskov, A., Astakhov, S. A., and Grassberger, P., "Least-dependent-component analysis based on mutual information", *Physical review E*, Vol. 70, No. 6, 2004.

[177] Benachir, D., Hosseini, S., Deville, Y., Karoui, M. S., and Hameurlain, A., "Modified independent component analysis for initializing non-negative matrix factorization: An approach to hyperspectral image unmixing", in *Electronics, Control, Measurement, Signals and their application to Mechatronics (ECMSM), 2013 IEEE 11th International Workshop of*, 2013, pp. 1-6.

Résumé en français

1. Contexte et objectifs

Tout au long du développement d'un produit pharmaceutique, il est important de contrôler la qualité des échantillons fabriqués. En effet, afin de garantir l'effet du produit sur le patient, et pour assurer la santé de ce dernier, le médicament se doit de répondre aux exigences réglementaires décrites dans les dossiers soumis aux agences [3]. Pour cela, de nombreuses méthodes analytiques sont disponibles permettant, par exemple, de doser le principe actif ou d'étudier sa libération dans l'organisme. La plupart de ces méthodes sont basées sur des analyses chimiques longues, qui détruisent l'échantillon, qui nécessitent l'utilisation de solvants, et sont consommatrices en ressources humaines et matérielles.

Depuis plusieurs années, les techniques de spectroscopies vibrationnelles, telles que la spectroscopie de diffusion Raman, ont fait leur apparition dans les laboratoires de contrôle [4]. Ces techniques ont montré de nombreux avantages car elles permettent en général des mesures rapides, sans solvants et sans destruction de l'échantillon. L'information contenue dans un spectre permet par exemple, de quantifier un principe actif dans un comprimé, d'authentifier un produit suspecté d'être falsifié ou bien de contrôler l'apparition d'une nouvelle forme cristalline au cœur du produit.

L'apparition des systèmes d'imagerie chimique a permis d'ajouter une nouvelle dimension spatiale en plus de la dimension spectrale classiquement utilisée [14]. En effet, avec ces techniques, les spectres sont acquis pour chaque pixel d'une image hyperspectrale, permettant d'obtenir une information sur la répartition des composés au sein d'un comprimé ou d'une poudre (Figure R-1). Par conséquent, lors du développement du médicament ou après sa mise sur le marché, l'étude de la distribution en actifs et excipients apporte une information complémentaire aux analystes pour assurer la qualité du produit ou pour comprendre un problème (exemple : modification du profil de dissolution à cause d'une mauvaise répartition d'un des produits).

Cependant, les techniques d'imagerie hyperspectrale fournissent des volumes de données importants qui ne sont en général pas interprétables par analyse visuelle et directe. C'est pourquoi les algorithmes chimiométriques apparaissent comme des outils incontournables pour extraire l'information pertinente de ces données [15]. De nombreuses méthodes sont

disponibles, allant des méthodes exploratoires non supervisées aux méthodes quantitatives supervisées. L'étude de la distribution des composés, l'identification des agglomérats au cœur d'une formulation, l'étude des formes cristallines dans un produit ont ainsi pu être étudiés en faisant appel à ces techniques.

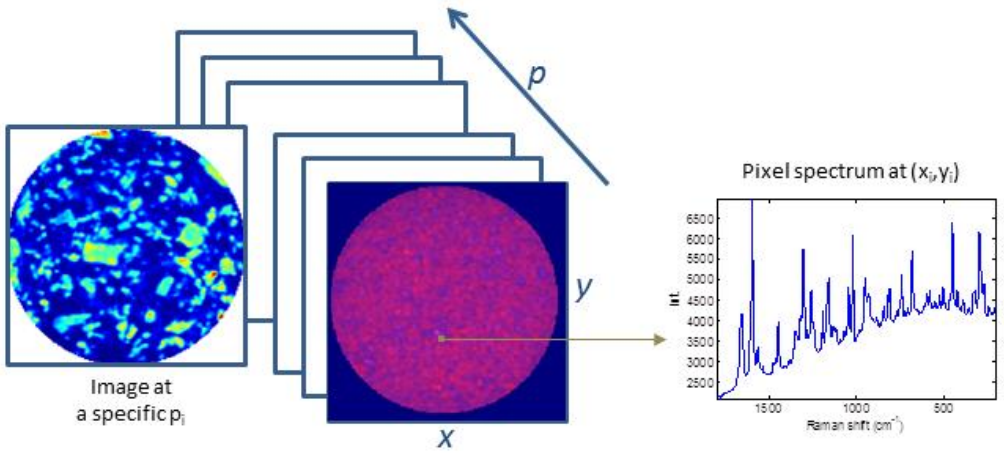


Figure R-1 Représentation schématique d'une image hyperspectrale Raman

Alors que la plupart des algorithmes classiquement utilisés sont basés sur des notions de variance spectrale, ou plus généralement sur la décomposition de moments statistiques, certaines limitations peuvent être rencontrées dans le cas d'un composé faiblement dosé. Dans cette thèse, un composé faiblement dosé est défini comme un composé ayant de faibles contributions spatiales et spectrales au sein de l'hypercube de données. En d'autres termes, ce composé est distribué de façon hétérogène au sein de l'échantillon analysé, c'est-à-dire qu'il est présent dans quelques pixels de l'image seulement. De plus, l'information spatiale portée par ce composé étant faible, elle peut être soit mélangée avec celle des autres produits, soit dispersée dans l'information non structurée associée au bruit.

Seul le cas du composé ayant de faibles contributions spatiales et spectrales sera étudié dans ces travaux (Table R-1). Les résultats sont obtenus sur des données issues de problématiques pharmaceutiques, mais les conclusions et approches proposées pourront être étendues au cas général d'un composé minoritaire dans un hypercube de données constitué d'un mélange de signaux.

		Contribution spatiale	
		Dans une majorité de pixels	**Dans quelques Pixels seulement**
Contribution spectrale	**Contributions spectrales fortes**	Détermination simple des spectres purs et cartes de distribution	Détermination simple des spectres purs et cartes de distribution
	Contributions spectrales faibles	Faible contribution spectrale du composé sur l'ensemble de l'image	**Contributions spectrales et spatiales faibles**

Table R-1 Contributions spatiales et spectrales d'un composé

A partir de données acquises à l'aide d'un microscope Raman, les objectifs de cette thèse peuvent être divisés en deux grands axes :

1/ Etudier la distribution en actifs et excipients dans une forme pharmaceutique solide

2/ Rechercher un constituant minoritaire d'une formulation dans une image de mélange

Cette thèse est organisée en 7 chapitres. Le premier chapitre présente les objectifs et le plan du manuscrit. Le deuxième chapitre présente le contexte de la thèse et effectue un état de l'art des outils et techniques utilisées dans ces travaux. La spectroscopie Raman et l'imagerie hyperspectrale, qui sont les techniques utilisées pour l'analyse des comprimés, sont décrites succinctement et un état des lieux des applications dans l'industrie pharmaceutique est présenté. Puis, sont présentés les prétraitements et outils chimiométriques utilisés pour les différentes études. De plus, l'application à la recherche d'un composé faiblement dosé, et la présentation de la problématique sont discutées. Les chapitres III à VI sont la reproduction de publications publiées ou soumises. Ces 4 publications sont introduites par une partie « préambule », puis conclues par une partie « contributions ». Dans les chapitres III et IV, les algorithmes d'analyse en composantes indépendantes (ICA) et de résolution multivariée de courbes par moindres carrés alternés (MCR-ALS) sont utilisés pour étudier la répartition des produits au sein d'une forme pharmaceutique solide. Les algorithmes sont challengés pour identifier un composé présent en faible quantité dans le comprimé. Les limites de ces approches sont mises en évidence pour la résolution de ce cas précis et des propositions de travail sont avancées. Les chapitres V et VI se focaliseront sur l'information spectrale uniquement, en favorisant l'espace des signaux, qui semble plus adapté pour la recherche d'un composé

minoritaire. Le chapitre V se concentre sur l'optimisation d'une contrainte spatiale pour la résolution du système. Cette méthode se base sur les projections orthogonales et semble tout à fait adaptée à l'identification d'un composé faiblement dosé. Le chapitre VI propose une méthode de détection des spectres purs d'un mélange en supposant que la composition de celui-ci n'est pas connue a priori. Cette méthode est testée avec succès sur une formulation qui contient des produits minoritaires.

2. Matériel et méthodes

2.1. Instrumentation et échantillons

Les études portent sur des images Raman de comprimés pharmaceutiques composés d'un ou plusieurs principes actifs et de différents excipients. Toutefois, Les méthodologies et approches testées pourront être appliquées à d'autres signaux ou échantillons.

Les images sont acquises en utilisant la cartographie Raman, qui permet d'enregistrer de façon séquentielle les spectres de l'image en déplaçant l'échantillon entre chaque acquisition spectrale. Après analyses des données, les images acquises permettent de caractériser et d'étudier la distribution des produits au sein de l'échantillon analysé.

Chaque pixel contient un spectre Raman qui peut être associé à un spectre de mélange des différents constituants du produit étudié. En fonction de la concentration du produit et de sa réponse spectrale, sa contribution au sein du spectre du mélange sera plus ou moins importante. De plus, le mélange des signaux sera directement dépendant de la résolution spatiale du système. En effet, plus la résolution spatiale sera faible (supérieure à la centaine de microns par exemple), plus l'information acquise pour chacun des pixels contiendra un mélange des produits de la formulation. A l'inverse, une résolution spatiale élevée (quelques microns par exemple), aura tendance à limiter les mélanges de signaux, au détriment de temps d'acquisition beaucoup plus longs. Afin d'avoir une surface représentative du comprimé, il est donc important de sélectionner une résolution spatiale en accord avec les objectifs de l'étude.

2.2. Analyse des données

Les différentes étapes d'analyse d'un hypercube de données suivent en général le déroulement suivant [48]:

- Déplier l'image pour obtenir une image 2-dimensions

- Prétraiter les données (corriger des variations indésirables ou exacerber de faibles modifications spectrales)
- Utiliser des algorithmes qualitatifs ou quantitatifs, supervisés ou non supervisés
- Réorganiser les résultats pour obtenir des cartes de distribution
- Appliquer un traitement d'image pour améliorer la qualité des résultats obtenus

Dans ces travaux, l'objectif sera de challenger deux approches chimiométriques, l'ICA et la MCR-ALS pour étudier leurs capacités à extraire l'information des constituants d'une forme pharmaceutique solide, incluant l'information portée par un composé faiblement dosé.

3. Contributions

L'identification et l'étude de la distribution d'un composé faiblement dosé peuvent être considérées comme le fil conducteur des travaux de cette thèse. Un composé faiblement dosé est considéré comme étant un produit présent dans quelques pixels de l'image avec une contribution spectrale faible. En utilisant les approches chimiométriques classiquement appliquées sur ce type de données, l'étude de la distribution d'un tel composé dans une image apparait comme un véritable challenge.

En effet, alors que les algorithmes chimiométriques se basent principalement sur la décomposition de moments statistiques, ces approches pourraient s'avérer limitées pour extraire l'information d'un constituant ayant de faibles contributions spatiales et spectrales. Dans la première partie de la thèse, constituée des chapitres III et IV, cette hypothèse est confirmée et les limitations des algorithmes ICA et MCR-ALS sont mises en évidence. Différentes propositions sont fournies afin d'aller chercher l'information liée au composé minoritaire. Bien que différentes pour les deux algorithmes testés, elles nécessitent d'aller regarder plus « profondément » dans la donnée car l'information du constituant minoritaire se situe principalement dans une partie moins structurée du signal.

Les parties suivantes de la thèse, constituées des chapitres V et VI, se focalisent sur un référentiel de travail différent, basé uniquement sur un espace spectral, décrivant un espace à P-dimensions (un axe par variable p) dans lequel les spectres peuvent être représentés comme des vecteurs. En utilisant cet espace, il est alors possible, grâce aux projections orthogonales, de s'affranchir progressivement des contributions spectrales d'un produit voire d'identifier des signaux dans une matrice. Ce nouveau paradigme permet dans le chapitre V de déterminer des cartes d'absence/présence à utiliser comme contrainte lors du processus itératif de la MCR-ALS. Dans le chapitre VI, l'utilisation d'une bibliothèque spectrale et de cet espace permet de

proposer une méthodologie innovante pour l'identification des produits purs dans une formulation inconnue.

4. Résultats

4.1. Utilisation de la séparation de source aveugle pour la détermination de spectres purs et l'étude de la distribution spatiale des composés

L'analyse en composantes indépendantes [80] est utilisée pour extraire des signaux purs à partir d'une image hyperspectrale Raman de comprimé. En théorie, cette approche recherche l'indépendance statistique des signaux dans un jeu de données et ne nécessite pas de connaissance à priori sur la formulation. Plusieurs algorithmes peuvent être utilisés pour effectuer une décomposition ICA. Dans nos travaux, l'algorithme JADE (Joint Approximate Diagonalization of Eigenmatrices) [123], basé sur l'optimisation des cumulants d'ordres 2 et 4, est utilisé. Cette algorithme ne nécessite pas de recherche de gradient et évite les problèmes de convergences qui peuvent être rencontrés avec d'autres algorithmes. L'algorithme décompose une matrice de spectre **X** comme suit :

$$\mathbf{X} = \mathbf{AS} \qquad \textbf{(R-1)}$$

Avec **S** une matrice des *k* sources indépendantes et **A** la matrice de mixage des coefficients.

Dans un premier temps, le travail consiste à démontrer les capacités de cet algorithme pour extraire l'information spectrale des constituants purs d'une formulation. Dans un second temps, l'importance de la sélection du nombre de composantes pour la décomposition est mise en évidence. En effet, le nombre de composantes (ICs) peut être choisi en fonction de la connaissance du produit (nombre de composantes égal au nombre de composés de la formulation) ou en utilisant une approche mathématique (ICA_by_blocs [124]). La sélection du nombre d'ICs apparait comme étant un élément critique de la décomposition. En effet, en utilisant un nombre d'ICs trop petit, les signaux extraits risquent d'être des signaux de mélanges de produits de l'échantillon analysé. A l'inverse, un nombre d'ICs trop grand risque de décomposer un signal en plusieurs contributions, voire à extraire des composantes de bruit. Afin d'obtenir un modèle et des signaux de qualité, l'étude montre qu'il est donc fondamental de sélectionner un nombre de composantes approprié. Pour cela, la méthode ICA_by_blocs apparait comme étant un bon compromis (Figure R-2). En utilisant cette approche, le nombre de composantes utilisé reste supérieur au nombre de constituants du mélange. Toutefois, celui-ci se rapproche de la réalité physico-chimique du mélange, incluant la variabilité chimique (différents

composés dans un mélange) et la variabilité physique des constituants (différentes tailles granulométriques, formes cristallines…).

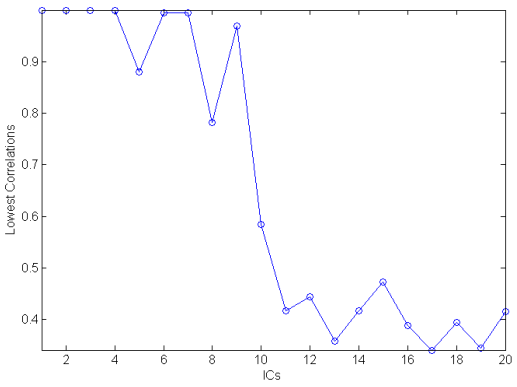


Figure R-2 Résultat d'ICA_by_blocs : Corrélations les plus faibles entre les signaux des deux blocs

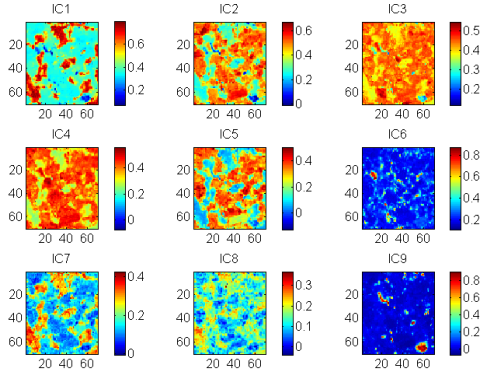


Figure R-3 Carte de distribution des coefficients A pour les 9 ICs

Dans la formulation étudiée, l'approche ICA_by_blocs nous a dirigé vers un nombre de composantes égal à 9, ce qui a permis d'obtenir les cartes de distributions associées aux 9 signaux calculés (Figure R-3). Alors que certaines cartes sont liées à un seul composé (IC6 et IC9), d'autres sont liées à un même produit (exemple IC2, IC3, IC4 et IC5 pour le lactose). Cette décomposition s'explique par la variabilité physico-chimique propre à un composé.

Pure spectrum	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	IC9
API1	0.01	-0.09	0.07	0.14	-0.04	0.13	0.21	0.18	**0.92**
API2	0.06	-0.01	0.11	0.08	0.03	**0.96**	0.08	0.10	-0.06
Lactose	0.25	**0.44**	**0.23**	**0.25**	**0.47**	0.00	0.36	0.45	-0.17
Avicel	**0.49**	0.15	0.06	0.02	0.20	-0.07	**0.38**	**0.61**	-0.20
Magnesium Stearate	0.20	0.00	0.01	0.04	0.04	0.41	0.32	0.23	-0.12

Table R-1 Corrélations entre les signaux du modèle et les spectres purs des constituants de la formulation

En utilisant cette approche, l'information liée au composé faiblement dosé n'est pas extraite (Table R-1). En effet, de par sa faible contribution (spatiale et spectrale), l'information associée est masquée par les autres constituants de la formulation et est contenue dans une part non expliquée du modèle. Afin d'identifier ce composé faiblement dosé, un modèle est construit avec

un nombre de composantes supérieur au nombre de constituants de la formulation et au résultat fourni par l'approche ICA_by_blocs. L'étude des corrélations entre les signaux calculés et le spectre pur du composé minoritaire permet d'identifier ce constituant sur la composante 12 d'un modèle ICA à 15 composantes (Figure R-4 et Figure R-5).

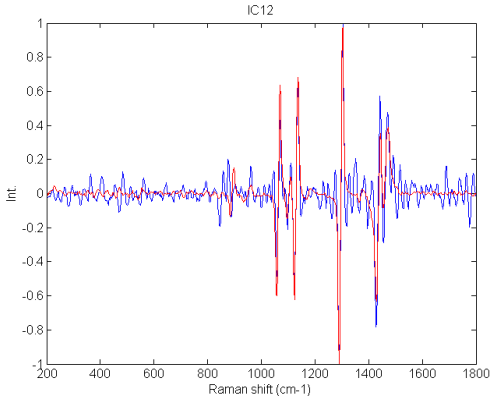


Figure R-4 Spectre dérivé du stéarate de magnésium (rouge) et IC12 (bleu)

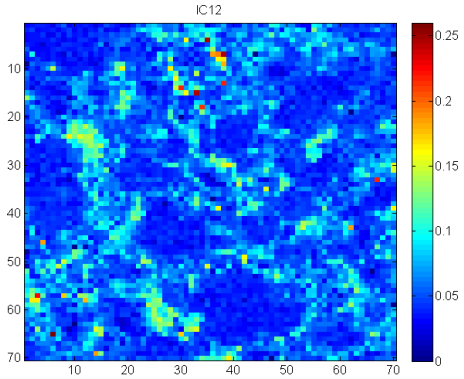


Figure R-5 Distribution du stéarate de magnésium dans l'image

4.2. Utilisation de résolution multivariée de courbes pour l'identification d'un constituant faiblement dosé

L'algorithme MCR-ALS [82] est utilisé pour étudier la répartition d'un composé faiblement dosé au sein d'une forme pharmaceutique solide. Déjà utilisée sur des images hyperspectrales Raman, l'approche MCR-ALS pour la détection d'un composé minoritaire reste toutefois un vrai challenge.

Cet algorithme fait appel à des contraintes appliquées sur les spectres ou les concentrations à chaque itération, ce qui permet de limiter les ambiguïtés (rotationnelles ou d'intensités) et de tendre vers une solution unique.

L'utilisation classique de l'algorithme MCR-ALS débute par un filtrage des données, équivalent à une réduction de dimensions qui utilise la décomposition en valeurs singulières de la matrice initiale, en utilisant un nombre de composantes *k* égal au nombre de constituants de la formulation tel que :

$$\mathbf{D_{PCA_{(n,p)}} = U_{(n,k)} S_{(k,k)} V^{T}_{(k,p)}} \qquad \textbf{(R-2)}$$

Cette réduction de dimension permet de réduire l'espace de travail en conservant l'information dite utile et en réduisant la part de bruit contenue dans l'hypercube de données.

Dans cette partie, il est démontré que la réduction de dimensions, dans le cas d'un composé faiblement dosé, doit être effectuée avec parcimonie ou totalement évitée afin de ne pas perdre l'information du composé minoritaire. En effet, dans ce cas, la part de variance spectrale portée par le produit est faible et ne se retrouve pas dans les premières composantes d'un modèle (si la matrice $\mathbf{D_{PCA_{(n,p)}}}$ est construite avec un nombre *k* trop faible) mais plutôt dans la part du bruit non-structuré.

Nombre de composantes *k* pour le calcul de $\mathbf{D_{PCA_{(n,p)}}}$	**5**	**10**	**15**	**20**	**50**
Iterations	9	5	5	3	3
R^2	99.4	98.8	98.7	98.6	98.4
Lof (%)	7.9	11.12	11.6	11.9	12.8
Cor. $S_{opt}1$/API1	0.98	0.98	0.98	0.98	0.98
Cor. $S_{opt}2$/API2	0.97	0.97	0.97	0.97	0.97
Cor. $S_{opt}3$/lactose	0.99	0.99	0.99	0.99	0.99
Cor. $S_{opt}4$/cellulose	0.95	0.95	0.95	0.95	0.95
Cor. $S_{opt}5$/MgSt	0.08	-0.01	-0.02	**0.87**	**0.90**

Table R-2 Résultats de la MCR-ALS avec un nombre de composantes *k* croissant pour la construction de la matrice $\mathbf{D_{PCA_{(n,p)}}}$

Les résultats des études sont présentés dans le tableau R-2. Pour un nombre de composantes *k* inférieur à 20, l'information liée au stéarate de magnésium n'est pas identifiée. En effet, avec un nombre de composantes inférieur à 20, l'information associée au composé minoritaire n'est pas inclue dans la matrice $\mathbf{D_{PCA_{(n,p)}}}$ et est contenue dans la part de variance non-expliquée du modèle. Pour assurer l'extraction de la contribution du constituant minoritaire, il est donc nécessaire de s'assurer d'avoir conservée l'information dans la matrice $\mathbf{D_{PCA_{(n,p)}}}$.

Pour cela, deux possibilités sont proposées :

- La matrice initiale $\mathbf{D_{PCA_{(n,p)}}}$ est construite avec un nombre suffisant de composantes *k*. Il y aura plus de bruit dans les signaux mais l'information du composé minoritaire sera conservée.

- La matrice initiale est « augmentée » en ajoutant des spectres purs liés au constituant minoritaire. La matrice $\mathbf{D_{PCA_{(n,p)}}}$, conservera donc cette information lors de la réduction de dimensions.

Dans ces deux cas, le prétraitement des données et l'utilisation des contraintes sont des paramètres critiques de la résolution. Les résultats obtenus permettent d'afficher la répartition des 2 actifs et des 3 excipients de la formulation étudiée, incluant le composé minoritaire (C_{opt5} sur la Figure R-6).

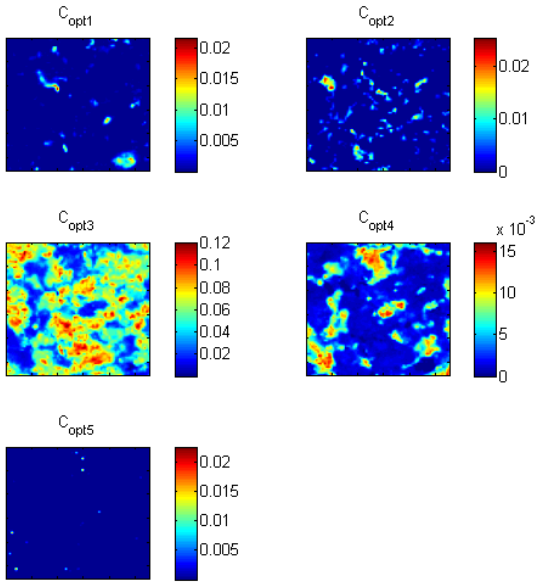


Figure R-6 Distributions des 5 constituants de la formulation (Calculées à partir d'une matrice non filtrée)

La contrainte d'égalité, basée sur la détermination des rangs locaux, et permettant, pour chaque pixel de déterminer l'absence ou la présence d'un constituant, s'avère indispensable pour obtenir des résultats satisfaisants. Toutefois, l'approche classique utilise des décompositions en valeurs singulières locales et la détermination d'un seuil [84]. Sans connaissance à priori de la distribution des constituants de l'image utilisée, la détermination de ces cartes pourrait s'avérer difficile. Une approche alternative est donc proposée dans le chapitre suivant.

4.3. Proposition d'une méthode pour la mise au point des cartographies d'absence/présence de composés

Les contraintes d'égalité, dont l'efficacité a été démontrée pour améliorer les résultats de l'algorithme MCR-ALS, peuvent parfois être difficiles à définir, notamment dans le cas d'un constituant faiblement dosé car le mode de calcul ne permet pas d'identifier simplement un composé avec de faibles contributions spatiales et spectrales. La méthode usuelle, basée sur l'approche FSIW-EFA, peut être décomposée de façon simplifiée par les trois étapes ci-dessous [144] :

- Effectuer une décomposition en valeurs singulières sur des blocs de pixels d'une image.
- Etudier la répartition de ces valeurs ordonnées dans un ordre croissant. L'utilisation d'un seuil permettra de définir le nombre de composés dans un bloc, et donc de déterminer le nombre de produits absents pour chaque bloc.
- Calculer les corrélations avec les spectres purs connus. L'absence d'un composé pourra être mise en évidence et associée à une valeur de 0 (ou à une valeur faible) lors du processus itératif de la MCR-ALS.

Dans le cas d'un constituant faiblement dosé, l'information spectrale du composé est souvent mélangée et masquée par les contributions spectrales des autres produits de la formulation. Les étapes de décomposition en valeurs singulières ou de corrélations avec les spectres purs peuvent montrer certaines limites dans ce cas précis.

L'approche alternative proposée se base sur un concept différent, qui fait appel uniquement à l'espace des signaux et peut donc être parfaitement adaptée aux composés minoritaires. Cette méthode utilise les projections orthogonales. Chaque constituant a son propre espace de variabilité défini par un ensemble de vecteurs incluant variabilités chimiques et physiques. Celui-ci est déterminé en faisant une acquisition d'image du produit puis en réalisant une décomposition en valeurs singulières non centrée sur l'image dépliée. Le nombre de vecteurs propres est sélectionné en observant la variance expliquée, qui doit être supérieure à 99.9% de la variance totale de l'image.

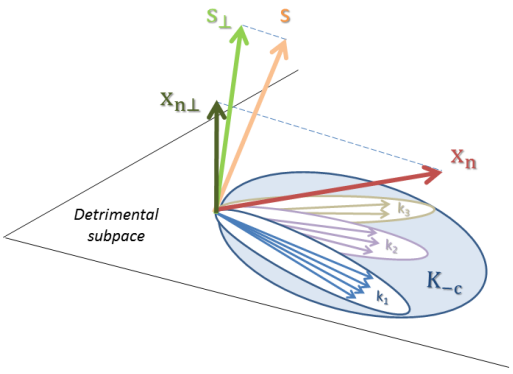


Figure R-7 Représentation graphique de l'approche proposée

Pour chaque constituant *c* de la formulation, une matrice d'interférence $\mathbf{K_{-c}}$ est déterminée. Cette matrice intègre la variabilité de tous les composés de la formulation, sauf la variabilité de *c*. Chaque spectre $\mathbf{x_n}$ (spectre de l'image) ou $\mathbf{s}$ (spectre de référence) est projeté orthogonalement à la base $\mathbf{K_{-c}}$ afin de ne conserver uniquement l'information utile de *c*. Pour cela, un projecteur orthogonal $\mathbf{\Sigma_{-c}}$ est calculé tel que :

$$\mathbf{\Sigma_{-c}} = \mathbf{I} - \mathbf{K_{-c}^T}(\mathbf{K_{-c}}\mathbf{K_{-c}^T})^{-1}\mathbf{K_{-c}} \qquad \textbf{(R-3)}$$

Les spectres sont projetés orthogonalement à la base $\mathbf{K_{-c}}$ en appliquant (Figure R-7) :

$$\mathbf{x_{n\perp}} = \mathbf{x_n}\mathbf{\Sigma_{-c}} \qquad \textbf{(R-4)}$$

et

$$\mathbf{s_\perp} = \mathbf{s}\mathbf{\Sigma_{-c}} \qquad \textbf{(R-5)}$$

L'étude des corrélations (équation R-6) entre les spectres projetés de l'image $\mathbf{x_{n\perp}}$ et les spectres purs projetés $\mathbf{s_\perp}$ détermine des cartes d'absence/présence à utiliser dans le processus itératif de la MCR-ALS.

$$\mathbf{r_{(x_{n\perp},s_\perp)_i}} = \frac{1}{\mathbf{n-1}} \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_{n\perp i}} - \overline{\mathbf{x_{n\perp}}})(\mathbf{s_\perp} - \overline{\mathbf{s_\perp}})}{\mathbf{S_{x_{n\perp i}} S_{s_\perp}}} \qquad \textbf{(R-6)}$$

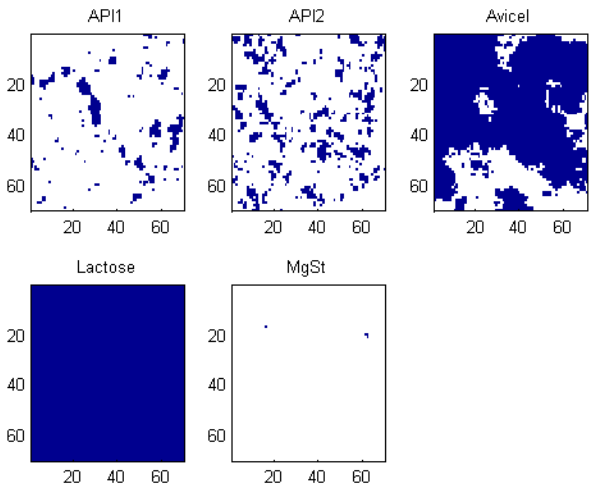


Figure R-8 Carte d'absence (blanc) / présence (bleu)

Dans l'exemple étudié, les cartes d'absence/présence (Figure R-8) définies par l'approche qui utilise les projections orthogonales ont permis d'améliorer considérablement les résultats de la MCR-ALS. En effet, la distribution du composé minoritaire au sein du comprimé a ainsi pu être fournie avec succès (Table R-3).

	MCR-ALS with n-n constraints	**MCR-ALS with n-n and local rank constraints**
$\mathbf{r_{sopt1/API1}}$	0. 70	0. 93
$\mathbf{r_{sopt2/API2}}$	0. 52	0. 93
$\mathbf{r_{sopt3/Avicel®}}$	0. 95	0. 81
$\mathbf{r_{sopt4/Lactose}}$	0. 98	0. 99
$\mathbf{r_{sopt5/MgSt}}$	0.34	0.84
lof %	2.66	10.09
Explained variance %	99.92	98.98

Table R-3 Résultats de la MCR-ALS

L'utilisation d'un espace spectrale permet, pour chaque produit, de s'affranchir de l'information des autres composés du mélange et facilite la mise au point des cartes d'absence/présence.

4.4. Approche itérative pour la détection des composés d'une formulation inconnue

Dans les précédentes parties de cette thèse, les formulations étudiées sont supposées connues. Toutefois, certaines applications ne permettent pas de connaitre les constituants d'un produit a

priori. C'est le cas, par exemple, des produits contrefaits (différents principes actifs ou excipients) ou de l'analyse d'un produit au cours d'une étude de stabilité (dégradation du principe actif, modification des formes cristallines…). Dans ces cas spécifiques, une approche de détection des composés s'avèrent très utile.

L'approche proposée dans ces travaux se base sur une bibliothèque spectrale, des calculs de distances, et des projections orthogonales. En travaillant exclusivement dans un espace des signaux, cette approche est parfaitement adaptée aux constituants faiblement dosés dans une formulation.

Soit $\mathbf{X}$ l'image hyperspectrale dépliée, $\mathbf{R_c}$ l'image dépliée d'un produit pur de la bibliothèque spectrale, et $\overline{\mathbf{R}}$ une matrice qui contient les spectres moyens de chaque image $\mathbf{R_c}$ de la bibliothèque.

L'approche proposée se divise en 5 étapes listées ci-dessous et résumées dans la Figure R-9 :

1. Initialiser le processus avec $\mathbf{X}_\perp = \mathbf{X}$ et $\overline{\mathbf{R}}_\perp = \overline{\mathbf{R}}$
2. Calculer les distances spectrales (SAM values) entre chaque spectre de $\mathbf{X}_\perp$ et chaque spectre moyen $\overline{\mathbf{R}}$
3. En observant les distances, identifier un produit pur *i* de la formulation et ajouter le spectre moyen associé dans la matrice **S**. Les itérations s'arrêtent si aucune faible valeur de distance n'est observée pour tous les produits de la bibliothèque spectrale.
4. Les spectres $\mathbf{X}$ et les spectres moyens $\overline{\mathbf{R}}$ sont projetés orthogonalement à l'espace vectoriel **K** constitué par les vecteurs propres déterminés par une décomposition en valeurs singulières non centrée sur les images dépliées des produits purs identifiés
5. Retour à l'étape 2

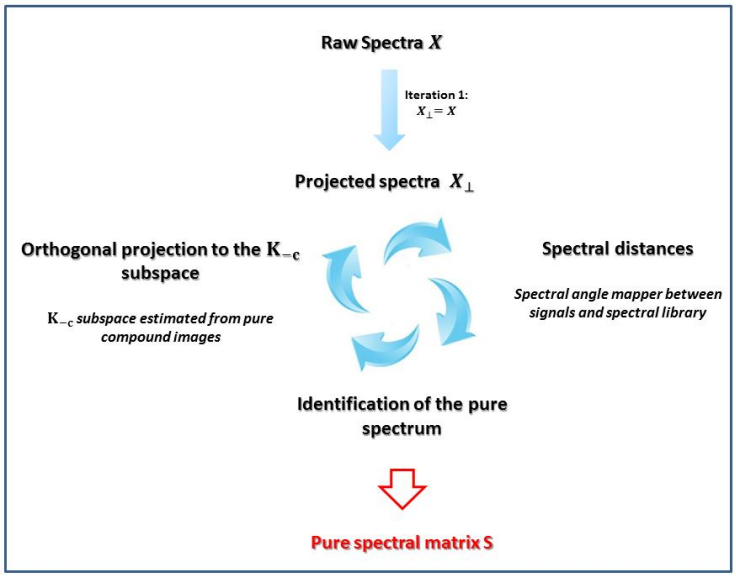


Figure R-9 Processus itératif de la méthode

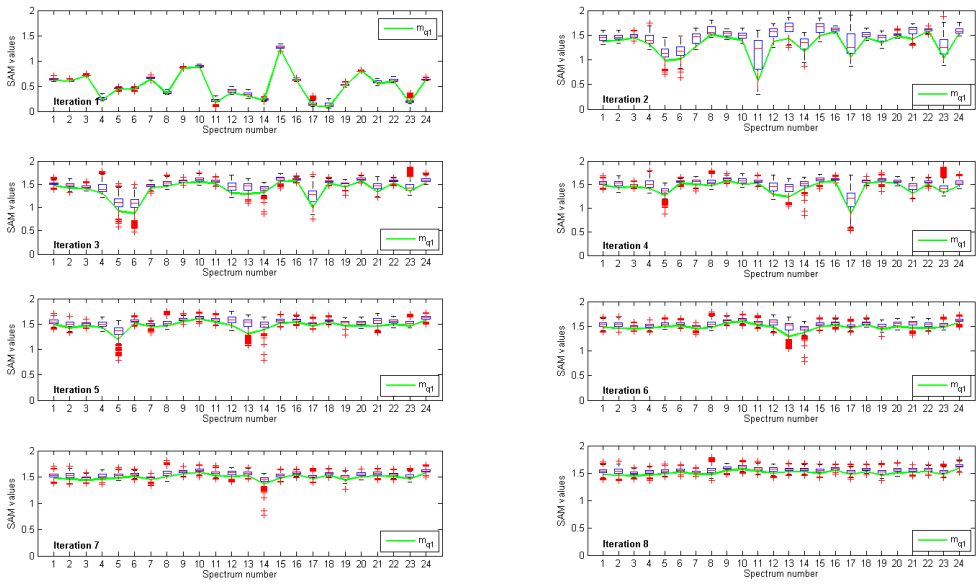


Figure R-10 Distances spectrales de l'iteration 1 à l'iteration 8

La Figure R-10 fournit pour 8 itérations, la distribution des distances spectrales calculées entre les spectres et les références. Des valeurs faibles sont associées à une similarité entre les deux signaux, et donc à la présence d'un composé. Des constituants sont ainsi identifiés jusqu'à l'itération 7, en mettant en évidence des valeurs moyennes ou des individus avec des valeurs de distance faibles. Chaque spectre identifié sera ajouté dans une matrice **S** utilisée pour initialiser la MCR-ALS qui fournira des valeurs proches de la formulation théorique (Table R-3).

o

Pure compound	Theoretical amount (% w/w)	Calculated C_c (% w/w)
Metolose®	40	39
Eudragit®	25	29
API	11	5 (form 1) + 5 (form 2)
Microcrystalline cellulose	17	15
Maltodextrin	6.5	6
Magnesium stearate	0.5	1

Table R-3 Résultats de la MCR-ALS à partir de la matrice S

5. Conclusions

Dans cette thèse, la microscopie Raman a été utilisée sur des formes pharmaceutiques solides pour étudier la répartition en actifs et excipients au cœur de ces échantillons. La détection d'un composé faiblement dosé a été le principal challenge de ces travaux. En effet, alors que la majeure partie des algorithmes se base sur des décompositions de moments statistiques, les difficultés d'extraction d'une information faible, distribuée dans quelques pixels de l'image et mixée avec les signaux des autres composés, ont rapidement été mises en évidence.

Dans ces travaux, l'ICA et la MCR-ALS sont utilisés avec succès pour étudier la distribution des composés dans un comprimé, sous réserve que ceux-ci soient présents avec une contribution suffisante au sein de l'image hyperspectrale. Dans le cas d'un composé faiblement dosé, les limitations de ces algorithmes ont été confirmées. Principalement liée aux modes de calcul, ou aux étapes de filtrage utilisées dans ces algorithmes, la détection d'un composé faiblement dosé montre de nombreuses difficultés.

Afin de pallier ces difficultés, différentes propositions de travail sont étudiées. Dans la première partie de ces travaux, l'ICA et la MCR-ALS sont utilisés. Dans le cas de l'ICA, un nombre élevé de composantes indépendantes s'avère nécessaire pour l'étude d'un composé minoritaire. En effet, en utilisant un nombre réduit de composantes (sélectionné manuellement ou avec un méthode statistique), l'information du composé minoritaire, qui se situe majoritairement dans une part de variance non-expliquée, ne se retrouve pas dans le processus de calcul et est donc perdue dès l'initialisation de l'algorithme. Un problème identique est observé en faisant appel à l'algorithme MCR-ALS ou la première étape du calcul consiste à réduire les dimensions du jeu de données.

Pour contourner ces limitations, deux méthodologies de travail sont proposées pour détecter un composé minoritaire. Dans le cas de l'ICA, l'utilisation d'un grand nombre de composantes sera nécessaire pour détecter un signal corrélé au constituant d'intérêt. Toutefois, en appliquant cette

méthodologie, il faut s'attendre à une diminution de la qualité des signaux pour les autres composés de la formulation. Dans le cas de la MCR-ALS, il est nécessaire de conserver un maximum d'information spectrale avant que le processus itératif ne démarre. Pour cela, la réduction de dimension initiale peut être contournée ou effectuée avec un nombre de composantes suffisant, ou l'utilisation des matrices augmentées pourra être envisagée. Afin d'obtenir une résolution satisfaisante, les prétraitements et les contraintes devront être optimisés.

Dans la seconde partie de la thèse, les travaux se focalisent sur l'espace des signaux, qui permet de s'affranchir d'une variabilité entre les individus, qui peut être très faible dans le cas d'un composé minoritaire.

Une première proposition consiste à développer de nouvelles cartes d'absence/présence d'un constituant, utilisées comme contraintes dans le processus de la MCR-ALS. Cette approche est une alternative aux méthodes actuelles basées à la fois sur des décompositions en valeurs singulières par blocs de pixels et sur des corrélations entre spectres. Pour chaque constituant, la matrice de spectres de l'image est projetée orthogonalement à une base d'interférence qui est constituée de toute la variabilité physico-chimique des composés autres que celui d'intérêt. En comparant les spectres résiduels avec les spectres de référence projetés, il devient alors possible de déterminer les cartes d'absence/présence pour un composé, incluant le cas d'un produit minoritaire.

La seconde proposition utilise également l'espace de signaux, mais cette fois ci pour identifier les constituants d'une formulation sans connaissance a priori, en s'appuyant uniquement sur une bibliothèque spectrale, des distances spectrales et des projections orthogonales. Cet algorithme itératif identifie pas à pas les composés d'une formulation.

Ces travaux démontrent donc les limitations des approches classiques, qui utilisent la décomposition de moment statistiques, pour l'identification des composés minoritaires dans un mélange. Dans ce cas précis, l'utilisation de l'espace des signaux est privilégié. Cette approche apporte une vraie valeur ajoutée à l'analyse des données dans le cas d'un composé minoritaire ayant de faibles contributions spectrales et spatiales dans une matrice de mélange.

Résumé

L'imagerie hyperspectrale est désormais considérée comme un outil analytique à part entière dans l'industrie pharmaceutique, aussi bien au cours du développement pour assurer la qualité d'un produit que pour résoudre des problématiques de production après la mise sur le marché du médicament.

Dans ces travaux, la microscopie Raman est utilisée pour étudier la distribution en principes actifs et excipients au sein d'une forme pharmaceutique solide, en se focalisant tout particulièrement sur l'identification d'un composé faiblement dosé. Ce dernier est défini comme étant un produit ayant de faibles contributions spatiales et spectrales, signifiant qu'il est distribué dans quelques pixels de l'image avec une information spectrale peu présente dans un spectre de mélange. Alors que la plupart des algorithmes chimiométriques se basent sur la décomposition de moments statistiques, nécessitant une variation suffisante entre les échantillons (les pixels d'une image), les limites de ces outils pour résoudre ce cas spécifique sont rapidement atteintes.

La première partie de la thèse met en évidence les difficultés de détection d'un composé faiblement dosé en utilisant l'analyse en composantes indépendantes et la résolution multivariée de courbes. Des méthodologies de travail sont proposées pour contourner ces limitations. Pour les deux techniques, les étapes de réduction de dimensions apparaissent comme des paramètres critiques de la méthode.

La seconde partie de la thèse se focalise sur l'espace des signaux pour déterminer des cartes d'absence/présence de constituants ou pour détecter des constituants dans une formulation inconnue, en se basant sur des espaces spectraux portant une information relative aux constituants de la formulation. Les techniques proposées sont parfaitement adaptées à la détection d'un composé faiblement dosé et ces méthodes pourraient être adaptées à d'autres techniques de mesure ou d'autres domaines d'application.

Mots clés : Microscopie Raman, constituant faiblement dosé, analyse en composantes indépendantes, résolution multivariée de courbes, projections orthogonales

Abstract

Hyperspectral imaging is now considered as a powerful analytical tool in the pharmaceutical environment, both during development to ensure the drug product quality and to solve production issues on commercialized products.

In this thesis, Raman microscopy is used to study the distribution of actives and excipients in a pharmaceutical drug product, by especially focusing on the identification of a low dose compound. This latter product is defined as a compound which has low spatial and spectral contributions, meaning that it is scattered in a few pixels of the image and that its spectral response is mixed with the other compounds of the formulation. While most chemometric tools are based on the decomposition of statistical moments (requiring sufficient variations between samples or image pixels), some limitations have been rapidly reached.

The first part of this thesis highlights the difficulty to detect a low dose compound in a product by using independent component analysis or multivariate curve resolution. Different methodologies are proposed to circumvent these limitations. For both techniques, reduction of dimensions and filtering steps appears as critical parameters of the method.

The second part of the thesis focusses on the signal space to determine absence/presence compound maps or to detect pure compounds in an unknown formulation. The proposed methods are only based on the spectral space of each formulation compound. There are perfectly suitable to a low dose compound and should be well-adapted to other analytical techniques or to other environments.

Keywords: Raman microscopy, low dose compound, independent component analysis, multivariate curve resolution, Orthogonal projections